Laboratoire d'Hydrologie
et Modélisation
Université Montpellier II

Laboratoire d'Hydrologie
ORSTOM Montpellier

ETUDE D'EVALUATION DE LOGICIELS D'AJUSTEMENTS DE LOIS STATISTIQUES SUR DES VARIABLES HYDROLOGIQUES

H. LUBES

J.M. MASSON

Février 1992

SOMMAIRE

1. INTRODUCTION

L'étude des risques associés à un événement hydrologique (crue, sécheresse...) passe toujours par l'analyse statistique d'observations. L'utilisation de formules empiriques régionales ou d'abaques n'échappe pas à cette règle : pour les établir, il a fallu procéder à l'analyse statistique d'observations.

Rappelons que l'analyse statistique des observations du passé a pour objectif d'obtenir des informations sur la population d'où elles sont tirées, afin de pouvoir énoncer des probabilités concernant l'avenir. La démarche statistique, qui suppose la stabilité de la population, s'effectue en trois étapes :

- Sélection d'observations en rapport avec le phénomène étudié. L'échantillon ainsi constitué doit avoir certaines qualités pour qu'on puisse en tirer des informations concernant la population.

- Ajustement d'une loi de probabilité théorique à la distribution de fréquence de cet échantillon. Il existe un éventail très large de lois théoriques et de méthodes d'ajustement de ces lois.

- Utilisation des résultats de l'étape précédente pour énoncer des probabilités concernant l'avenir.

Les logiciels étudiés ont pour but de faciliter la réalisation de la deuxième étape de la démarche, et laissent à l'hydrologue la responsabilité des étapes précédentes et suivantes. Cependant, même pour cette seconde étape, compte tenu du fait qu'aucune théorie ne permet de choisir la loi suivie par la plupart des variables hydrologiques, et que les tests d'adéquation des ajustements donnent seulement des indications et non des certitudes, le jugement de l'hydrologue est préservé et les logiciels n'aboutissent en aucun cas au choix automatique d'une loi : ils laissent toujours l'hydrologue décider. Ainsi, grâce aux logiciels, l'hydrologue a plus de temps et plus d'informations pour exercer son jugement, ce qui ne simplifie pas forcément sa tâche. En effet, une loi de probabilité est un modèle qui doit être à la fois <u>descriptif</u> et <u>prédictif</u>.

D'une part ce modèle doit bien s'ajuster à l'échantillon des observations, et d'autre part il doit correctement estimer les risques, c'est-à-dire la probabilité des événements dans la population.

Même si on élimine les erreurs diverses qui peuvent affecter l'échantillon des observations (erreurs de mesure, hétérogénéité c'est-à-dire mélange de populations,...), du seul fait d'un tirage aléatoire, un échantillon peut présenter des particularités : sous représentation ou sur représentation des fortes ou des faibles valeurs par exemple, et le risque est d'autant plus grand que l'échantillon est petit. Autrement dit, un échantillon représente plus ou moins bien la population.

Bien ajuster le modèle (loi de probabilité) à l'échantillon est facile. Il suffit de choisir des lois à 3 ou 4 paramètres - associées à une transformation de la variable (logarithme ou racine carrée par exemple). Mais la fonction de répartition qui grâce à ces nombreux paramètres va parfaitement décrire les sinuosités de la distribution de l'échantillon, risque de conduire à des extrapolations éloignées de la réalité du fait des courbures qui permettent à cette fonction de répartition de passer par les valeurs extrêmes observées.

Un moyen d'augmenter l'information apportée sur une station de mesure par un petit échantillon, est d'utiliser des observations de stations voisines qui ont un comportement identique, autrement dit de procéder à une analyse régionale.

Pour une analyse régionale, les lois à 3 ou 4 paramètres présentent un inconvénient majeur lié aux choix des méthodes d'ajustement. Si des méthodes d'ajustement différentes (moments, maximum de vraisemblance...) appliquées sur le même échantillon conduisent à des valeurs voisines en ce qui concerne les quantiles, elles peuvent conduire à des jeux de paramètres très différents dont la combinaison donne cependant des fonctions de répartition très voisines. Il en va ainsi des lois log-Pearson et gamma généralisées (Bobée et Ashkar, 1991). Si on veut étudier régionalement la valeur des paramètres, mieux vaut donc prendre des lois plus simples ou des caractéristiques statistiques plus stables (moments).

Un bon modèle prédictif est un modèle robuste, peu sensible aux fluctuations d'échantillonnage. Il donne des résultats voisins avec des échantillons qui ont des particularités différentes. Les lois de probabilité comportant peu de paramètres sont les plus robustes, mais les logiciels de ce point de vue n'apportent pas d'aide significative, aucun d'eux ne mettant en oeuvre des tests ou des indices de robustesse.

2. LES PRODUITS EVALUES

Le présent rapport rassemble les conclusions d'une étude d'évaluation de logiciels d'ajustements de lois statistiques sur des variables hydrologiques.

Cette étude a consisté à décrire les fonctionnalités de chaque logiciel, à étudier l'exactitude du contenu statistique, et à formuler des appréciations.

Les avis apportés débouchent sur des propositions pour concevoir un nouveau logiciel dont la particularité serait de guider l'utilisateur tout au long de sa démarche statistique.

Les logiciels étudiés fonctionnent sur micro-ordinateurs. Nous avons comparé les produits suivants classés par ordre alphabétique :

ALED, Logiciel développé en QUICK BASIC par le Laboratoire d'Hydrologie et Modélisation (L.H.M.) de l'Université Montpellier 2.

DIXLOI, Logiciel développé pour l'essentiel en FORTRAN 77 par le Laboratoire d'Hydrologie de l'ORSTOM (Montpellier).

HFA (Hydrologic Frequency Analysis), Logiciel développé par B. Bobée (Université de Québec, Canada) et F. Ashkar (Université de Moncton, Canada) pour l'analyse fréquentielle des événements extrêmes.

TROPHEE (Traitement des Observations Pluviométriques et Hydrométriques des Evénements Extrêmes), Logiciel développé par le BCEOM Société Française d'Ingéniérie (Grande-Motte, France), et dont le module "Traitements statistiques sur fichier hors base de données" a été mis à notre disposition.

Nous disposions par ailleurs de :

LOIS développé au CEMAGREF (Groupement d'Antony), et

CANTIL programme d'ajustement de lois du CEMAGREF accompagnant le document "Hydrologie appliquée aux petits bassins ruraux" de C. Michel (1989).

Nous ne faisons pas état ici des comparaisons effectuées avec LOIS (Mercier 1991). En effet LOIS n'a pas été totalement opérationnel sur le micro-ordinateur utilisé.

Il en est de même pour CANTIL qui est un petit programme d'application d'un chapitre d'un livre d'hydrologie ; il ne concerne que quelques lois, ne fait aucune sortie graphique, et la justification des expressions utilisées pour le calcul des intervalles de confiance pose encore quelques problèmes.

3. LES ECHANTILLONS D'OBSERVATIONS : SAISIE, TESTS EVENTUELS DE LEUR QUALITE, ET CARACTERISTIQUES PRINCIPALES

Pour les quatre logiciels étudiés, les échantillons traités sont :

- soit saisis au clavier
- soit contenus dans un fichier dont le format est imposé par le logiciel. Le plus souvent ces fichiers sont créés à l'issu d'une saisie au clavier.

3.1. SAISIE AU CLAVIER

Il faut distinguer le contenu de la saisie qui ne se limite pas aux seules valeurs numériques sur lesquelles portent les traitements statistiques, et que nous désignerons par observations, et les modalités de la saisie qui déterminent la plus ou moins grande facilité avec laquelle l'utilisateur crée son échantillon d'observations.

3.1.1 Contenu de la saisie

Le tableau I ci-après indique pour chaque logiciel les données à saisir.

Tableau I

LOGICIELS	Nom de l'échantillon (titre) 1	Identificateur de station 2	Type de coordonnées géographiques par lesquelles la station est repérée 3	Coordonnées géographiques de la station 4	Altitude de la station 5	Type de données (valeurs maximales...) 6	Nombre d'observations 7	Unité des observations 8	Effectif de chaque observation 9	Période d'observations et/ou nombre d'années d'observations 10	Date de référence de chaque observation 11	Nom du fichier où sont stockées les données 12	Code de fichier 13
ALED	Facultatif						X	Facultatif	X			Facultatif	
DIXLOI	Facultatif	Facultatif	Facultatif	Facultatif	Facultatif			Facultatif			Facultatif	X	X
HFA	Facultatif										Facultatif	X	
TROPHEE		Facultatif				X				X		X	

Remarque : Le terme "facultatif" désigne des données qui sont demandées à l'utilisateur (soit sous forme d'une question, soit d'un champ à compléter) mais dont la réponse est optionnelle, dans le sens où un simple retour chariot permet d'ignorer la demande sans que la validation ultérieure de la saisie soit remise en cause.

Les "X" désignent des données pour lesquelles l'utilisateur doit obligatoirement apporter une réponse.

Commentaires relatifs à chacune de ces données.

Certaines remarques peuvent être retenues comme des propositions ou des points de réflexion en vue de la réalisation d'un nouveau logiciel.
Le nombre entre parenthèses renvoie à la colonne du tableau I.

(1) il est souhaitable d'avoir la possibilité de désigner l'échantillon par un titre qui apparaît à l'impression ou à la visualisation des résultats ou graphiques issus des traitements statistiques. Cette donnée doit être facultative.

(2) l'identificateur de station (nom ou numéro) figure à l'impression ou à la visualisation des graphiques et résultats statistiques issus des traitements statistiques. Cette donnée doit être facultative.

(3) (4) (5) le repérage géographique et l'altitude de la station concernée par les observations ne s'imposent pas. En effet ces données ne sont pas toujours facilement disponibles, sauf si les observations sont issues d'une banque de données, mais les logiciels ne procèdent pas eux-mêmes à l'extraction de ces informations.

(6) ces précisions sont requises par TROPHEE. Les observations traitées par TROPHEE sont des valeurs maximales. L'utilisateur doit préciser s'il s'agit des maxima annuels ou de valeurs supérieures à un seuil. Le choix de l'une ou l'autre de ces caractéristiques permet de fixer ou non le paramètre de position des lois théoriques à 3 paramètres à la valeur du seuil. Ce choix est lié à la définition des conditions de l'ajustement, il semble donc préférable de le faire au moment de l'ajustement.

Par ailleurs, TROPHEE distingue deux types de données observées : valeurs instantanées ou valeurs moyennes sur des durées variables, puisqu'un module d'élaboration de courbes intensité-durée-fréquence (pour des données de pluie), a été développé.

(7) il peut être utile de rentrer le nombre des observations qui constituent l'échantillon pour tester si ce nombre est supérieur à une limite inférieure en deçà de laquelle il n'est pas raisonnable de procéder à un ajustement statistique. Un test analogue peut-être fait par rapport à un effectif maximum d'observations (200 dans ALED, 500 dans DIXLOI).

(8) l'unité des observations figure à l'impression ou la visualisation des graphiques et résultats statistiques issus des traitements statistiques. Cette donnée doit être facultative.

(9) la possibilité d'introduire des données groupées c'est-à-dire d'indiquer le nombre de fois où la même valeur de la variable a été observée est particulièrement intéressante, notamment pour l'étude des durées des épisodes pluvieux à partir des relevés pluviométriques : on observe généralement plusieurs dizaines d'épisodes de même durée exprimée en jours.

(10) la période que recouvre les observations est une donnée essentielle pour le calcul de la période de retour d'un événement. Toutefois il serait préférable que cette information soit demandée non pas à la constitution de l'échantillon mais à l'étape propre au calcul des périodes de retour.

(11) la possibilité d'indiquer une date de référence pour chaque observation est intéressante. Le repérage des observations saisies peut en être facilité. De plus cette référence peut être un critère de constitution des sous-échantillons permettant la mise en oeuvre du test de Mann-Whitney (Cf § 3.3.2). A ce titre la référence du mois peut être particulièrement utile.

(12) à l'exception de ALED, les logiciels imposent la sauvegarde de l'échantillon d'observations dans un fichier avant de procéder aux traitements statistiques. Ceci n'est pas une contrainte dans la mesure où il est toujours possible de détruire un fichier qu'il n'est pas utile de conserver sous le système d'exploitation.

(13) DIXLOI impose un code à tout fichier de données. Cette codification est spécifique au Laboratoire d'Hydrologie de l'ORSTOM qui développe des standards pour les fichiers de données, parmi lesquels les fichiers dits de type 21. Il n'est pas utile de prévoir ainsi une codification particulière propre à chaque utilisateur, une totale liberté existant au niveau du nom de fichier lui-même.

Compléments

- DIXLOI permet de saisir plusieurs échantillons dans un même fichier. Cette possibilité est intéressante pour un travail à la chaîne sur les observations d'un grand nombre de stations d'une même région ou d'un même pays par exemple.

- ALED et HFA affectent un numéro d'ordre aux observations, ce qui peut être une commodité de saisie.

- Les logiciels ont défini un nombre maximal d'observations constituant un échantillon : 200 valeurs au plus pour ALED et HFA, 500 valeurs pour DIXLOI, 200 pour TROPHEE.
  Il est nécessaire d'essayer de s'affranchir de cette contrainte, l'analyse statistique d'échantillons supérieurs aux limites données doit être possible notamment dans le cadre d'études régionales.

- En ce qui concerne la valeur numérique X de l'observation elle-même,

  pour DIXLOI, $X < 10^4$
  pour HFA, $0<X<10^6$
  pour TROPHEE, $X < 10^4$
  pour ALED, il n'y a pas de contrainte.

La justification d'une valeur limite supérieure est discutable.

- Les données hydrologiques traitées sont nulles ou strictement positives.

Les conditions d'ajustement d'échantillons comprenant des valeurs nulles et l'interprétation des résultats obtenus constituent un sujet délicat. L'exclusion des valeurs nulles dans de tels échantillons implique en toute rigueur d'en tenir compte dans la méthode d'ajustement, ce qui n'est pas simple. (Cf, § 11.1 et 12).

Seul HFA définit rigoureusement sa position sur le sujet en interdisant à la saisie une observation nulle, et en procédant à un ajustement classique ignorant l'existence éventuelle d'observations nulles non prises en compte dans l'échantillon.

ALED et DIXLOI acceptent un échantillon contenant des valeurs nulles à la saisie, et le traitent statistiquement à moins qu'une impossibilité mathématique (fonction logarithme...) ne survienne.

TROPHEE accepte à la saisie une observation nulle mais la remplace par la valeur minimale non nulle de l'échantillon pour procéder aux calculs statistiques sans que l'utilisateur en soit informé.

- DIXLOI, HFA et TROPHEE imposent un nombre maximal de chiffres après le point décimal (DIXLOI 2, HFA 2, TROPHEE 3) qui résulte vraisemblablement de la définition des grilles de saisie.

3.1.2. Modalités de saisie

ALED

La saisie a lieu ligne à ligne.
Les réponses erronées sont signalées (le message "vous n'avez pas tapé un caractère correct" apparaît), et doivent être corrigées pour progresser dans le déroulement du logiciel.
En fin de saisie les observations peuvent être revisualisées à l'écran ou sur imprimante.
Il est possible ensuite d'apporter des corrections sous réserve d'avoir repéré le nombre d'observations à corriger et leur numéro d'ordre.
Au niveau de la phase de saisie, il n'est pas possible de supprimer une observation ou d'en insérer une nouvelle.
Toutefois, la structure du fichier de sauvegarde des observations, particulièrement simple, permet facilement ces opérations au moyen d'un éditeur ASCII (Cf § 3.2).
Il n'est pas possible enfin d'interrompre la saisie en sauvegardant les données rentrées.

DIXLOI

DIXLOI utilise un masque de saisie et dispose donc des fonctionnalités qui lui sont attachées : définition de champs numériques, alphanumériques...
Une première grille de saisie définit les caractéristiques générales du fichier (titre, code, nombre d'échantillons, présence de dates ou non) et donc communes à tous les échantillons le constituant.

Deux types de grille de saisie des valeurs numériques sont ensuite proposés en fonction de la présence ou non de dates de référence des observations.
Huit caractères numériques sont réservés pour les dates sans signification imposée, donc aucun contrôle de validité n'est effectué.
Chaque grille permet la saisie plein écran de 100 valeurs. Des corrections éventuelles peuvent être apportées grille par grille à l'aide de touches de fonction, le retour à la grille précédente étant impossible.
L'insertion d'une donnée sur une grille incomplète au sein d'autres observations est impossible, elle ne peut avoir lieu qu'en séquence après la dernière valeur saisie.
La saisie terminée, il n'est plus possible sans quitter le logiciel de revisualiser les données et a fortiori d'apporter des corrections avant de procéder au traitement statistique. Ce qui est regrettable étant donné la structure rigide du fichier de sauvegarde des observations (Cf 3.2) dont la manipulation sous éditeur ASCII nécessite beaucoup d'attention.
Il n'est pas possible de reprendre une saisie interrompue.

HFA

HFA utilise un masque de saisie et ses attributs. Les observations sont introduites et revisualisées une à une. Chaque observation est affectée d'un numéro d'ordre.
Des touches de fonction sont prédéfinies et notamment une fonction "HELP" qui spécifie les conditions de saisie.
Les dates, si elle sont rentrées, doivent respecter le schéma YYYY-MM-DD (année-mois-jour). Des contrôles sont effectuées sur leur validité.
Les touches ↑ ↓ permettent de parcourir à tout instant le fichier, observation après observation. Les touches Del et Ins permettent de supprimer une observation ou d'en insérer une autre à n'importe quel endroit du fichier.
Après sauvegarde de l'échantillon celui-ci peut-être immédiatement réédité sur la même grille d'écran pour contrôles complémentaires.

On peut regretter qu'à aucun moment une visualisation plein écran ne soit possible, mais la structure simple du fichier de sauvegarde rend ce mode de "lecture" aisé sous éditeur ASCII. De plus une option d'impression des données est opérationnelle.
Des sauvegardes en cours de saisie peuvent être exécutées ce qui est une possibilité appréciable.

TROPHEE

La saisie a lieu par l'intermédiaire d'un masque de saisie type "plein écran" où les caractéristiques des champs (numériques...) sont prédéterminées, et de fenêtres.
Des touches ↑↓ ← → permettent de parcourir les données. Les touches Del et Ins sont opérationnelles.
Après sauvegarde du fichier, il est possible en ne quittant pas cet environnement de reprendre immédiatement la saisie, ce qui peut se limiter à une visualisation de l'échantillon pour un dernier contrôle.

3.2. STRUCTURE DES FICHIERS DE DONNEES

Sous réserve de respecter la structure des fichiers ASCII créés après saisie au clavier, des échantillons peuvent être constitués à "l'extérieur" du logiciel. L'intérêt est évident lorsque les observations à étudier résident déjà sur un support magnétique : il s'agit alors de les mettre dans la forme et le type de codage ASCII attendus. Les logiciels ne procèdent pas eux-mêmes à cette réécriture qui nécessite donc un développement particulier.
Par ailleurs certains utilisateurs familiarisés à la manipulation d'un éditeur ASCII donné souhaitent constituer un échantillon par ce moyen.
Toutes ces opérations sont d'autant plus faciles que la structure imposée par le logiciel est simple. Cette simplicité est liée à la nature des données sauvegardées.

ALED stocke le titre de l'échantillon s'il existe et les valeurs observées avec éventuellement leur nombre d'occurences.

La structure est la suivante :

- soit 1ère ligne : titre (60 caractères)
  puis une donnée par ligne sans format imposé

- soit 1ère ligne : titre (60 caractères)
  puis par ligne : donnée "," nombre d'occurences (ex : 41,1)
  seul le séparateur "," est imposé.

Cette structure simple permet facilement d'apporter des modifications à l'aide d'un éditeur ASCII, ou de créer un tel fichier par programme informatique.

DIXLOI : la structure des fichiers de données est rigoureusement spécifiée par un ensemble de formats qui rendent relativement lourde une manipulation sous éditeur.

La structure des fichiers HFA est élémentaire.

1ère ligne : titre (caractères)
puis soit une donnée par ligne
soit par ligne : donnée et date de référence sans aucun séparateur (ex : 41.0019910124 pour 41 mm le 24 janvier 1991)

Il est simple de manipuler ce fichier sous éditeur ASCII ou de constituer un programme d'écriture pour le générer.

TROPHEE : le fichier comprend une première série de lignes où sont notées les informations du type : titre, type de données, période, seuil éventuel..., puis une deuxième série où figurent les valeurs numériques des observations.

Toutes ces données sont enregistrées en format libre. Il est donc possible de manipuler ou de générer un tel fichier à l'extérieur de TROPHEE sans trop de difficultés, sous réserve toutefois de s'informer auprès des concepteurs des conventions minimales à respecter.

3.3 VERIFICATION DE LA QUALITE DES ECHANTILLONS

3.3.1. Qualités des échantillons

Pour être représentatif de la population d'où il est tiré, l'échantillon des observations doit présenter un certain nombre de qualités qui ont été bien précisées par Bobée et Ashkar (1991). Les observations contenues dans l'échantillon doivent être :

-Aléatoires : c'est-à-dire, en hydrologie, être le résultat de fluctuations naturelles et non la conséquence d'influences anthropiques.

- Indépendantes : la valeur d'une observation ne doit pas être influencée par la valeur de l'observation précédente au sens chronologique. Les séries chronologiques de débits journaliers par exemple sont souvent aléatoires mais jamais indépendantes. Les logiciels ne sont pas prévus pour traiter les séries chronologiques de variables dépendantes.

- Homogènes : c'est-à-dire provenant d'une même population. Il est souvent difficile de trancher dans ce domaine : peut-on mélanger les crues pluviales et celles de fonte des neiges ? Les valeurs extrêmes telle la pluie de Nîmes le 3 octobre 1988 ne constituent-elles pas une population à part ? Certains pays (Italie) les traitent comme telles (Masson, 1992).

- Stationnaires : les variables appartenant à une série chronologique sont dites stationnaires quand leurs caractéristiques statistiques (moyenne, autocovariance) ne changent pas avec les saisons. En hydrologie, les variations naturelles dues aux saisons peuvent être neutralisées en découpant l'année en périodes pendant lesquelles on considère que la variable est stationnaire. On ne mélangera pas les observations mensuelles de janvier et de juillet. Cependant, d'une année à l'autre, indépendamment des fluctuations climatiques séculaires difficiles à mettre en évidence, il peut se produire des changements brutaux (jumps en anglais) suite à des aménagements ou des évolutions plus ou moins régulières (tendances) dues par exemple au changement d'occupation des terres (urbanisation).

3.3.2. Tests mis en oeuvre par les logiciels

Seul, le logiciel HFA propose des tests non paramétriques pour vérifier la qualité de l'échantillon des observations.

** TEST D'INDEPENDANCE ET DE TENDANCE*

Le test proposé est celui de Wald et Wolfowitz (1943). Ce test est peu souvent cité dans les ouvrages statistiques en français et nous ne l'avons rencontré que dans Lebart et Fenelon (1975) pour tester le caractère aléatoire de séquences d'observations d'une variable qualitative à deux modalités (jeu de pile ou face par exemple).

Tel qu'il est utilisé par le logiciel HFA, le test se présente ainsi :

x_1, x_2,......x_i, ...x_n sont les observations de l'échantillon dans leur succession chronologique.

On calcule :

$$R = \sum_{i=1}^{n-1} x_i \, x_{i+1} + x_1 \, x_n$$

Si les observations successives sont indépendantes, R suit une loi asymptotiquement normale de moyenne :

$$\bar{R} = \frac{(s_1^2 - s_2)}{(n-1)}$$

et de variance :

$$Var(R) = \frac{(s_2^2 - s_4)}{(n-1)} - \bar{R}^2 + \frac{s_1^4 - 4\, s_1^2\, s_2 + 4\, s_1\, s_3 + s_2^2 - 2\, s_4}{(n-1)\,(n-2)}$$

avec $s_r = n\, m'_r$, m'_r étant le moment d'ordre r par rapport à l'origine des observations.

On peut donc transformer en une variable normale réduite U_r la valeur de R calculée sur les observations, et rejeter l'indépendance si $|U_r| > U_{1-\alpha/2}$ en prenant un risque α (5 % par exemple) de se tromper si les observations successives sont vraiment indépendantes.

Il faudrait tester par simulation la capacité de ce test à détecter des tendances. D'après Kendall et Stuart (1943), ce test ne serait pas plus efficace qu'une régression linéaire.

* *TEST D'HOMOGENEITE ET DE STATIONNARITE*

Le second test proposé par le logiciel HFA est le test de Mann-Whitney qu'on trouve dans le chapitre des tests non paramétriques des ouvrages statistiques de base comme celui du C.E.R.E.S.T.A (1986) . Ce test permet de décider si 2 échantillons sont tirés ou non de la même population. Pour l'appliquer, il faut couper l'échantillon des observations en deux parties, la coupure étant par exemple liée à un changement d'appareil de mesure ou à un aménagement. S'agissant du débit maximal annuel d'une rivière, on peut aussi vérifier en fonction de la saison où il se produit, si on a affaire à la même population.

Dans HFA la définition des deux échantillons se fait de manière interactive à partir de deux types de graphiques possibles :

-soit la chronologie des valeurs observées avec les valeurs en ordonnée et le temps en abscisse (les données doivent être affectées au moins d'un identificateur année).

-soit les fréquences mensuelles des événements avec les mois en abscisse et les fréquences en ordonnée (les données doivent être affectées au moins d'un identificateur mois).

* *TEST DE DETECTION DES VALEURS EXTREMES INFERIEURES OU SUPERIEURES : Les horsains (outliers en anglais)* (Masson, 1992)

Ce test est sujet à discussions dans la mesure où des tirages aléatoires artificiels dans une population donnée fournissent des échantillons avec des valeurs extrêmes éliminées par les tests.
Le test proposé par le logiciel HFA est celui de Grubbs et Beck (1972) qui ne convient que si la population est normale. Pour tenir compte de cette contrainte, le logiciel HFA travaille sur le logarithme des observations, comme le recommande le Conseil des Ressources en eau des Etats Unis d'Amérique, ce qui ne constitue toutefois qu'une approximation quand la loi log-normale ne convient pas.

Les valeurs extrêmes inférieures et supérieures centrées réduites sont comparées à des valeurs tabulées pour le niveau de signification $\alpha = 0.1$. En fait, les valeurs tabulées ont été remplacées par une approximation polynomiale fonction de n, la taille de l'échantillon.

HFA donne une représentation graphique du test de Grubbs et Beck. L'axe des abscisses représente le numéro d'ordre des observations dans l'échantillon. Les valeurs observées sont portées en ordonnée. Les limites inférieures et supérieures estimées par le test définissant les zones de horsains, sont représentées par des lignes horizontales.

3.3.3. Propositions

Nous ne pensons pas qu'il soit nécessaire ni même utile de faire beaucoup de tests sur la qualité des échantillons qui est essentiellement sous la responsabilité de l'utilisateur.

- Pour les valeurs extrêmes :

  ou bien il s'agit d'erreurs de saisie et il suffit d'afficher le minimum et le maximum pour d'un seul coup d'oeil les détecter, ou bien la valeur a été réellement mesurée et il est difficile de justifier son élimination.

- Pour l'homogénéité et la stationnarité :

  Les variations naturelles dues aux saisons sont connues des hydrologues. Si un accident majeur a affecté la série, il est en général connu de tous (début de la sécheresse en 1968-70 au Sahel, mise en service du Barrage Seine en 1965...). Si cet accident est mineur, les tests non paramétriques, peu puissants, ont peu de chance de le détecter.
  Toutefois nous proposons de présenter les graphiques suivants qui permettent de détecter visuellement ces changements brutaux :
  -chronologie des valeurs observées avec les valeurs en ordonnée et le temps en abscisse,
  -fréquences mensuelles des événements avec les mois en abscisse et les fréquences en ordonnée.

- Indépendance :

  c'est une qualité qu'on peut envisager de tester, pour éliminer d'éventuelles séries chronologiques autocorrelées que des utilisateurs non avertis pourraient tenter de traiter avec le logiciel. Il suffirait de calculer le coefficient d'autocorrélation avec retard de 1 et afficher un message s'il se révélait être significativement différent de zéro au risque $\alpha = 0.1$ par exemple.

3.4. CALCUL DE QUELQUES CARACTERISTIQUES STATISTIQUES LIEES A L'ECHANTILLON

Il peut s'agir des caractéristiques statistiques de l'échantillon ou de celles de la population estimées à partir de l'échantillon après correction d'un biais éventuel.

Le logiciel ALED calcule :

- la valeur minimale de l'échantillon
- la valeur maximale de l'échantillon
- l'effectif de l'échantillon
- la moyenne arithmétique
- l'écart-type
- le coefficient d'asymétrie (ou de dissymétrie)
- le coefficient d'aplatissement

Les caractéristiques statistiques sont celles de la population estimées à partir de l'échantillon, après correction du biais éventuel.

Si x_i, i = 1, 2, ..., n sont les valeurs de l'échantillon, on a d'après Haan (1977) :

- moyenne arithmétique $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

- écart type : $\hat{\sigma}_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{(n-1)}}$

- coefficient de dissymétrie : $\hat{\gamma}_1 = \dfrac{n \sum_{i=1}^{n} (x_i-\bar{x})^3}{(n-1)\,(n-2)\,\hat{\sigma}_x^3}$

- coefficient d'aplatissement : $\hat{\gamma}_2 = \dfrac{n^2 \sum_{i=1}^{n} (x_i-\bar{x})^4}{(n-1)\,(n-2)\,(n-3)\,\hat{\sigma}_x^4}$

Par ailleurs ALED trace un histogramme des données. La largeur de la classe est estimée par l'algorithme proposé par D. W. Scott (1985).

Le logiciel DIXLOI fournit :

- la moyenne arithmétique
- la médiane observée
- le mode probable
- la variance et l'écart-type
- le coefficient de variation
- le coefficient d'asymétrie
- le coefficient d'aplatissement

Ces quatre dernières caractéristiques sont celles de l'échantillon.

Le logiciel HFA donne un tableau des moments de la population estimés à partir de l'échantillon. Une colonne concerne les valeurs brutes, l'autre colonne concerne les valeurs après transformation logarithmique. Ces moments sont :

- la moyenne arithmétique
- la moyenne harmonique
- la moyenne géométrique
- l'écart-type
- le coefficient de dissymétrie
- le coefficient de variation

Le logiciel TROPHEE fournit des caractéristiques de la population estimées à partir de l'échantillon :

- la moyenne arithmétique
- l'écart-type
- le coefficient de variation

Il est regrettable avec les logiciels DIXLOI, HFA et TROPHEE de ne pouvoir connaître ces informations qu'après avoir procédé à l'ajustement d'une loi théorique, puisque ces différentes caractéristiques qui ne sont pas visualisées à l'écran sont seulement rassemblées dans le fichier résultat généré après ajustement. Seul ALED par conséquent présente ces caractéristiques à

l'écran avant le menu des différentes lois théoriques traitées par le logiciel, ce qui peut guider l'utilisateur dans le choix de l'ajustement à réaliser.

Proposition

Puisqu'on s'intéresse à la population, il semble logique que ce soit ses caractéristiques statistiques qu'on cherche à estimer à partir de l'échantillon.

L'idée de traiter aussi les valeurs après transformation en logarithmes est intéressante . Elle permet d'avoir rapidement une idée de l'efficacité de cette transformation.

Il convient de se limiter à des caractéristiques calculables sans ambiguïté ; ce n'est pas le cas du mode, que nous éliminerons donc.
Sur les valeurs naturelles et transformées en logarithmes, on pourrait calculer

- la moyenne arithmétique
- la moyenne harmonique
- la moyenne géométrique
- la médiane
- l'écart-type et le coefficient de variation correspondant si la moyenne est significativement différente de zéro
- le coefficient de dissymétrie
- le coefficient d'aplatissement

Pour avoir une idée des dispersions d'échantillonnage, on pourrait calculer, comme le fait le logiciel STATGRAPHICS (Statistical graphics system) dans son menu "Statistiques Elémentaires", un certain nombre d'erreurs standards. L'erreur standard d'un estimateur sans biais est l'écart type de la distribution l'échantillonnage de cet estimateur.

- Erreur standard de la moyenne arithmétique x : $\hat{\sigma}_{\bar{x}} = \dfrac{\hat{\sigma}_x}{\sqrt{n}}$

qui permettrait de voir si la moyenne est significativement différente de zéro avant de calculer le coefficient de variation.

- Erreur standard des coefficients de dissymétrie et d'aplatissement dans le cas d'un distribution normale des variables :

$$\hat{\sigma}_{\hat{\gamma}_1} = \sqrt{\frac{6}{n}}$$

$$\hat{\sigma}_{\hat{\gamma}_2} = \sqrt{\frac{24}{n}}$$

Quand les variables suivent une loi normale on a aussi $\gamma_1 = 0$ et $\gamma_2 = 3$, et les estimateurs de ces quantités suivent une distribution normale autour de ces valeurs. On peut donc construire les variables normales réduites :

$$\frac{(\hat{\gamma}_1 - 0)\sqrt{n}}{\sqrt{6}} \quad \text{et} \quad \frac{(\hat{\gamma}_2 - 3)\sqrt{n}}{\sqrt{24}}$$

qui doivent être inférieures à 2 en valeur absolue dans 98 % des cas si la distribution des variables est normale.

- Le tracé d'un histogramme de fréquences des valeurs brutes paraît intéressant malgré le caractère subjectif de la répartition en classes, en tant que graphique de synthèse des principales caractéristiques de l'échantillon.

4. LES LOIS AJUSTEES

L'inventaire donne les résultats bruts suivants :

Présence (+) ou absence (-) dans le logiciel

Nom des Lois	ALED	DIXLOI	HFA	TROPHEE
Loi binomiale négative tronquée	+	-	-	-
Loi exponentielle 1 et 2 paramètres	+	-	-	+
Loi de Fréchet	-	+	-	+
Loi des fuites	+	+	-	-
Loi gamma incomplète à 2 paramètres et à 3 paramètres = Pearson III	+	+	+	+
Loi gamma généralisée	-	-	+	-
Loi géométrique	+	-	-	-
Loi de Gumbel	+	+	-	+
Loi de Jenkinson ou GEV	+	-	-	-
Loi log-gamma	-	+	+	-
Loi log-normale ou de Galton à 2 ou 3 paramètres	+	+	-	+
Loi normale	+	+	-	+
Loi de Pearson V	-	+	-	-
Loi de Polya	-	+	-	-
Loi de Weibull ou de Goodrich	+	+	-	+

Le logiciel HFA se limite à la famille des lois gamma. On peut théoriquement passer des lois gamma à presque toutes les autres lois. On trouve dans Bobée et Ashkar (1991) un tableau qui donne toutes les relations possibles avec les autres lois et les transformations de variables nécessaires pour y parvenir. Par exemple, de la loi gamma généralisée à 4 paramètres notée GG4(s, α, λ, m), on passe :

- à la loi de Pearson III en posant $s=1$
- à la loi de Weibull à 3 paramètres en posant $\lambda = 1$
- à la loi de Weibull à 2 paramètres en posant $\lambda = 1$ et $m = 0$

- à la loi de Gumbel à dissymétrie positive en posant $\lambda = 1$, $m = 0$ et en travaillant sur la variable $Y = -\text{Ln } X$.

Mais ces relations sont données pour les chercheurs et praticiens "qui souhaitent avoir un aperçu rapide de la manière dont les différentes distributions sont liées entre elles" et non comme un mode d'emploi pour ajuster les différentes lois grâce au logiciel puisque les auteurs montrent au chapitre 9 que deux lois : la log-gamma et la gamma généralisée, recouvrent l'ensemble des autres distributions qu'il est donc inutile d'utiliser.

A ce niveau d'abstraction statistique, nous pensons qu'il faut séparer théorie et pratique. En effet, les lois qui recouvrent la quasi totalité des autres lois ont 3 ou 4 paramètres et nous avons vu au § 1 que, selon les méthodes d'ajustement, ces paramètres peuvent prendre des valeurs très différentes tout en donnant des fonctions de répartition voisines. Comment, dans ces conditions, conclure valablement que tel paramètre se rapproche de la valeur 1 ou 0 et donc qu'on peut se satisfaire de telle ou telle loi plus simple ?

On constate d'autre part que les usages ont consacré l'adéquation de quelques lois simples à certaines variables hydrologiques : loi log-normale pour les débits mensuels, loi de Gumbel pour les hauteurs de pluie extrêmes sur une durée.... Pour ces variables, il est intéressant de disposer directement de la loi adéquate.

Les menus proposés par ALED et DIXLOI sont dans l'ensemble comparables. DIXLOI ajuste la loi log-gamma, la loi de Fréchet et la loi de Pearson V, ce que ne fait pas ALED qui propose par contre la loi de Jenkinson, la loi géométrique et la loi exponentielle.

DIXLOI traite les lois les plus usuelles. Deux d'entre elles ont des appellations qui sont peut-être à reprendre. La loi dite de Polya est en fait la loi binomiale négative. Quant à la loi de Goodrich, nous n'avons trouvé aucune référence bibliographique sur une loi de ce nom, hormis celle de Roche (1963) à l'origine de toutes les citations dans les ouvrages français.

TROPHEE propose les lois les plus communément utilisées en hydrologie.

Propositions

Nous n'avons trouvé aucune application hydrologique de la loi de Pearson V (loi gamma en 1/x) ; nous proposons donc de la supprimer. Il suffit de proposer la transformation de variable $y = \frac{1}{x}$ pour pouvoir quand même l'ajuster en cas de besoin.

La loi binomiale négative n'est pas utilisable directement pour décrire les durées des épisodes secs ou pluvieux. Cette loi donne en effet une certaine probabilité à la valeur zéro alors que par définition, à partir du moment où il existe, un épisode pluvieux ou sec ne peut avoir une durée nulle. Nous proposons donc de la remplacer par la loi binomiale négative tronquée (LBNT) qui s'applique à des variables entières strictement supérieures à zéro.

La panoplie des lois proposées dans un premier temps pourrait être :

- La loi normale
- La loi log-normale à 3 paramètres
- La loi gamma incomplète à 2 paramètres
- La loi gamma incomplète à 3 paramètres
- La loi exponentielle à 1 ou 2 paramètres
- La loi de Weibull à 2 ou 3 paramètres
- La loi de Gumbel
- La loi de Jenkinson
- La loi des fuites
- La loi log-gamma à 3 paramètres
- La loi géométrique
- La loi binomiale négative tronquée

Nous maintenons la loi de Jenkinson parce qu'elle est recommandée pour l'étude des crues au Royaume-Uni ainsi que la loi log-gamma qui joue le même rôle aux Etats-Unis d'Amérique.

La possibilité de transformer la variable origine X en une variable Y telle que :

$Y = Ln(X)$

permettrait d'ajuster directement :

- La loi log-normale à 2 paramètres en passant par la loi normale.

- La loi de Fréchet par l'intermédiaire de la loi de Gumbel (la loi de Fréchet est également accessible par la loi de Jenkinson)

- La loi log-gamma à 2 paramètres en passant par la loi gamma à 2 paramètres.

$Y = \frac{1}{X}$ permettrait d'ajuster la loi de Pearson V en passant par la loi gamma à 2 paramètres

$Y = \sqrt{X}$ donnerait satisfaction à ceux (certaines Agences de Bassin) qui ajustent sur des hauteurs précipitées mensuelles ou annuelles la loi racine carrée normale (Méthode du maximum de vraisemblance uniquement).

Pour cette première étape toutes les lois, sauf peut-être la loi de Weibull à 3 paramètres, existent dans l'un ou l'autre des logiciels. Il conviendrait toutefois de vérifier par un travail spécial (stagiaire par exemple) avant de les introduire dans le logiciel :

- La loi log-gamma à 3 paramètres (par comparaison avec les résultats de HFA).

- La loi de Weibull à 3 paramètres avec ajustement du paramètre de position.

Dans un second temps on pourrait ajouter d'autres lois, mais toujours après une étude vérifiant à la fois leur intérêt et les techniques de calcul. Sans aller jusqu'à étudier toutes les lois qui sont proposées dans la littérature et dont certaines ne semblent utilisées que par leurs auteurs comme Kumaraswamy (1980) ; ni sans remonter à des systèmes de lois d'origine française comme les lois de Halphen (Halphen 1949), on pourrait examiner :

- La loi gamma généralisée dont l'intérêt n'est pas évident à première vue et qu'on pourrait tester, au moins dans sa version à 3 paramètres, grâce au logiciel HFA.

- Une loi utilisée par E.D.F (Duband 1982) pour représenter la distribution des hauteurs de pluie sur les durées de 2 heures à 5 jours et dont la fonction de répartition :

$$F(x) = 1 - \alpha\, e^{-x/a} - \beta\, e^{-x/b}$$

donne une probabilité aux valeurs nulles puisque $F(0) = 1- \alpha - \beta$.

On pourrait d'ailleurs comparer cette fonction à un mélange de 2 distributions exponentielles dont la densité de probabilité serait :

$$f(x) = \alpha\, a\, e^{-ax} + (1-\alpha)\, b\, e^{-bx}$$

- De nouvelles lois de valeurs extrêmes proposées pour satisfaire aux conditions suivantes :

  - Bien reproduire la variabilité des observations
  - Etre peu sensibles aux valeurs extraordinaires (horsains)
  - Avoir une expression explicite
  - Comporter peu de paramètres et être faciles à calculer.

- Deux distributions semblent avoir beaucoup de ces qualités :

  - la distribution des valeurs extrêmes à deux composantes (Rossi, Fiorentino et Versace 1984).
  - la distribution Wakeby (Houghton 1978).

Cependant Wakeby a besoin de cinq paramètres.
La distribution à deux composantes n'a que 4 paramètres mais la méthode du maximum de vraisemblance qui les estime ne converge pas toujours, de plus, l'inversion de sa fonction de répartition n'est pas explicite.

D'après Ahmad, Sinclair et Werritty (1988), la loi log-logistique, qui ne comporte que 3 paramètres, aurait toutes les qualités requises et serait supérieure à toutes les lois de valeurs extrêmes utilisées jusqu'ici. Cela mériterait d'être vérifié !

5. METHODES D'ESTIMATION DES PARAMETRES DES LOIS

5.1. APERCU THEORIQUE

Les lois statistiques possèdent un ou plusieurs paramètres qui, dans la population d'où est tiré l'échantillon, ont une valeur donnée (vraie) mais inconnue. Il s'agit d'estimer cette valeur à partir des observations de l'échantillon.
Les estimations qu'on va faire donnent des résultats différents d'un échantillon à l'autre.

Le résultat des estimations est donc une variable aléatoire sur laquelle on peut calculer une moyenne, une variance...
Les résultats varient aussi avec les méthodes d'estimation. Selon les méthodes, les estimateurs n'auront pas la même moyenne, pas la même variance...

** RAPPEL SUR LES QUALITES DES ESTIMATEURS*

La qualité des estimateurs s'évalue. Si G est l'estimateur d'un paramètre γ, on dira que G est non biaisé si en moyenne on retrouve la vraie valeur du paramètre, c'est-à-dire si :

$$E[G] = \gamma$$

La précision d'un estimateur se mesure par son moment d'ordre 2 par rapport à la vraie valeur, c'est-à-dire par la quantité : $E[(G-\gamma)^2]$ qui est la variance de l'estimateur si celui ci est non biaisé. On démontre que cette précision ne peut-être inférieure à une valeur minimale (inégalité de Cramer-Rao).

Un estimateur non biaisé de variance minimale est un estimateur efficace,

l'efficacité d'un estimateur non biaisé étant le rapport de la variance minimale à sa propre variance. Pour tout paramètre, il existe au moins un estimateur asymptotiquement efficace (dont l'efficacité tend vers 1 quand l'effectif de l'échantillon tend vers l'infini).

Il existe d'autres propriétés que nous ne développerons pas car elles ne nous serviront pas.

** METHODES D'ESTIMATION*

On a d'une part une population où la variable X suit une loi théorique caractérisée par une expression mathématique : la fonction densité de probabilité qui comporte un ou plusieurs paramètres dont les vraies valeurs inconnues sont α, β, γ et que nous symboliserons par $f(x, \alpha, \beta, \gamma)$ et d'autre part un échantillon d'observations tirées de la population, c'est-à-dire une série de valeurs $x_1, x_2, \ldots x_i, \ldots x_n$.
Pour une valeur donnée de la variable : $X = x$, on a :

$$\text{Prob}(x < X \leq x + dx) = f(x, \alpha, \beta, \gamma)\, dx$$

** METHODE DU MAXIMUM DE VRAISEMBLANCE*

Compte tenu de la loi de probabilité supposée, la méthode maximise la fonction de vraisemblance L (α, β, γ), c'est-à-dire la probabilité d'obtenir les valeurs de l'échantillon des observations :

$$L(\alpha, \beta, \gamma) = \prod_{i=1}^{n} f(x_i, \alpha, \beta, \gamma)$$

Mathématiquement, on sait calculer les valeurs α, β, γ qui maximisent une fonction. Pour cela on annule les dérivées partielles de la fonction de vraisemblance par rapport à α, β, γ. Par commodité, on travaille sur le logarithme de la fonction de vraisemblance.

Avec 1 ou 2 paramètres, la méthode donne généralement une solution explicite. Avec 3 paramètres ou plus, on obtient un système d'équations à résoudre numériquement et la convergence n'est pas assurée dans tous les cas.

L'intérêt de la méthode est qu'elle fournit toujours :

- Des estimateurs de variance minimale ou asymptotiquement minimale, bien qu'ils ne soient pas toujours non biaisés, mais pour une loi donnée il est souvent possible de corriger le biais (Fiorentino et Gabriele 1984)

- Des estimateurs dont les distributions d'échantillonnage sont asymptotiquement normales.

** METHODE DES MOMENTS CLASSIQUES*

A partir de l'expression de la loi théorique, on peut exprimer les moments théoriques d'ordre k en fonction des paramètres de la loi. Ainsi, pour les moments par rapport à l'origine, on a :

$$m_k(\alpha, \beta, \gamma) = \int_{\text{domaine de X}} x^k f(x, \alpha, \beta, \gamma)\, dx$$

La méthode des moments égale les k premiers moments calculés sur l'échantillon des observations aux k expressions théoriques correspondantes (k = nombre de paramètres dans la loi).

La méthode donne généralement des relations explicites simples, ce qui a fait sa popularité. Les estimateurs obtenus par la méthode des moments ont des distributions d'échantillonnage asymptotiquement normales.

* *METHODES DIVERSES BASEES SUR LES MOMENTS*

Quand une loi utilise le logarithme de la variable Y = Ln X (loi log-normale, loi de Fréchet, loi log-gamma) on peut estimer les paramètres :

- soit à partir des moments de X et on parle alors de méthode des moments directe.
- soit à partir des moments de Y et on parle de méthode des moments indirecte.

Avec plus de 2 paramètres, la méthode des moments nécessite l'estimation des moments d'ordre supérieur à deux qui ont une très grande dispersion d'échantillonnage. Pour éviter de les utiliser, des auteurs ont proposé des méthodes mélangeant des moments ne dépassant pas l'ordre 2 (method of mixed moments = méthode des moments mélangés) utilisant par exemple la moyenne arithmétique, la variance et la moyenne géométrique et/ou la moyenne harmonique. Ces estimateurs ont les mêmes propriétés que les estimateurs de la méthode des moments classique avec peut-être une variance d'échantillonnage plus faible.

* *METHODE UTILISANT LE MODE et/ou LES QUANTILES*

Ces méthodes, développées surtout pour les lois de valeurs extrêmes, nécessitent qu'on associe à chaque valeur de l'échantillon, une fréquence de non dépassement (en utilisant une expression liée au rang r des valeurs classées (§6.1.1), par exemple (r-0,5)/n dont on reparlera plus loin).

Cette fréquence de non dépassement permet d'estimer, par interpolation sur l'échantillon, certains quantiles ou le mode quand celui-ci correspond à une valeur précise de la fonction de répartition comme dans la loi de Gumbel. Des expressions lient les paramètres de la loi aux valeurs du mode et des quantiles.

Ces méthodes donnent en général des estimateurs de qualité médiocre.

** METHODE DES MOINDRES CARRES*

Quand il existe une relation linéaire entre la variable et sa fonction de répartition, comme par exemple dans la loi de Gumbel :

$$x = x_0 + s\,[-\mathrm{Ln}\,(-\mathrm{Ln}\,F(x))]$$

les couples (x, F(x)), F(x) étant estimé par une expression liée au rang des valeurs classées, permettent à l'aide d'une régression linéaire classique, d'estimer les paramètres de la loi.

Du point de vue statistique, les estimateurs sont de qualité médiocre. Cependant Laborde (1984) a montré que cette méthode, appliquée sur la moitié supérieure des échantillons (rang n/2 à n) de pluies journalières maximales des stations pluviométriques lorraines, donnait des estimations du gradex meilleure que celles obtenues par la méthode des moments.

** METHODE BASEE SUR LES STATISTIQUES D'ORDRE*

Proposée par White (1964) pour la loi log-Weibull et donc utilisable pour la loi de Gumbel moyennant un changement de signe (Cf. § 4), cette méthode nécessite des tables de valeurs correspondant à chaque rang des observations classées pour chaque effectif d'échantillon, tables limitées à $n<20$. Bien que d'après Raynal et Salas (1986), elle donne de bons résultats, nous citons cette méthode pour mémoire.

** METHODE DU MAXIMUM D'ENTROPIE*

L'entropie d'un système complet de k événements E, mesure l'incertitude associée à la réalisation d'un événement.
Elle s'écrit :

$$H = -\sum_{J=1}^{k} P(E_J)\,.\,\mathrm{Ln}\left[P(E_J)\right]$$

$P(E_J)$ = probabilité de voir apparaître l'événement Ej au cours d'une épreuve.

On constate que l'entropie est maximale quand tous les événements ont la même probabilité d'apparaître.

A partir d'un échantillon d'observations :

$$x_1, x_2, ...x_i, ...x_n$$

On ne peut maximiser la fonction :

$$H = -\sum_{i=1}^{n} P(x_i) \, . \, Ln \left[P(x_i)\right]$$

que si on impose des contraintes. L'une de ces contraintes est générale : $\sum_{i=1}^{n} P(x_i) = 1.$

Pour les autres contraintes (autant que de paramètres dans la loi ajustée), on retient en général des espérances mathématiques qui dépendent donc de la loi ajustée. Leur expression générale est :

$$E\left[g_J(x)\right] = \sum_{i=1}^{n} g_J(x_i) \; P(x_i) \qquad J=1,m \; ; \; m = \text{nombre de paramètres.}$$

Avec une variable continue, P(x) est remplacé par f(x) dx, mais le principe est le même.

La méthode a surtout été développée par Jowitt (1979) pour la loi de Gumbel dont la fonction de répartition est $F(x) = e^{-e^{-(x-x_0)/s}}$. Les espérances mathématiques utilisées pour les contraintes sont dans ce cas :

$$E\left[\frac{x-x_0}{s}\right] = 0.5772 \text{ (constante d'Euler)}$$

$$E\left[e^{-\frac{x-x_0}{s}}\right] = 1$$

Pour la loi de Gumbel, des distributions d'échantillonnage ont été étudiées par Huynh (1986 - 1987). D'après cet auteur, si on considère à la fois le biais et la variance d'estimation, cette méthode est meilleure que toutes les autres.

Cependant il s'agit d'une méthode compliquée qui n'a pas été développée pour beaucoup de lois.

** METHODE DES MOMENTS DE PROBABILITE PONDERES*

Cette méthode a été introduite par Greenwood, Landwehr, Matalas et Wallis (1979) et s'est rapidement généralisée aux lois faciles à inverser. Par exemple, la loi de Gumbel, dont la fonction de répartition est $F(x) = e^{-e^{-(x-x_0)/s}}$ est facile à inverser. En effet, on peut facilement exprimer x en fonction de F (x). On obtient :

$$x = x_0 + s\,[-\mathrm{Ln}\,(-\mathrm{Ln}\,F(x))]$$

Sous leur forme générale, les moments de probabilité pondérés s'énoncent :

$$M_{ljk} = E\left[X^l F^j (1-F)^k\right] = \int_0^1 [x(F)]^l F^j (1-F)^k \, dF$$

Mais dans la plupart des applications pratiques, on utilise soit $M(k) = M_{10k}$, soit $M(j) = M_{1j0}$.

Lubès et Masson (1991) ont exposé l'utilisation de cette méthode à propos de la loi de Jenkinson.

Cette méthode est intéressante sous plusieurs aspects :

- La possibilité d'obtenir des estimateurs non biaisés des moments de probabilité pondérés
- Une distribution d'échantillonnage des estimateurs qui est asymptotiquement normale
- Une grande simplicité de mise en oeuvre quand la loi est facile à inverser.

Raynal et Salas (1986) recommandent cette méthode pour la loi de Gumbel et récemment, des auteurs chinois ont développé l'utilisation de cette méthode pour des lois difficiles à inverser : Song et Ding (1986), Jing, Song, Yang et Hou (1989).

5.2. METHODES PROPOSEES PAR LES LOGICIELS

DIXLOI et TROPHEE ne proposent que deux méthodes : maximum de vraisemblance et moments classiques.

ALED, propose aussi ces deux méthodes pour le plus grand nombre de lois, mais dans certains cas, impose d'autres méthodes. Ainsi pour la loi des fuites, la seule méthode proposée est une méthode des moments modifiée, tenant compte de la proportion de valeurs nulles. De même pour la loi de Jenkinson, la seule méthode proposée est la méthode des moments de probabilité pondérés.

HFA, qui ne traite que les lois gamma, propose par contre presque toutes les méthodes possibles d'ajustement (sauf la méthode des moments de probabilité pondérés).

5.3. PROPOSITIONS

L'objectif d'un logiciel d'ajustement de lois n'est pas de proposer toutes les méthodes d'ajustement, mais de fournir les meilleurs résultats. Il doit donc être un peu directif.

La méthode du maximum de vraisemblance qui, asymptotiquement, donne les estimateurs de variance minimale, devrait toujours être proposée. On lui associerait obligatoirement une autre méthode pour les raisons suivantes :

- Hosking et al. (1985) ont montré par simulation sur de petits échantillons que cette méthode peut donner des estimateurs plus variables que d'autres méthodes.

- Avec 3 paramètres ou plus, la méthode du maximum de vraisemblance conduit à une solution numérique par itération à partir de valeurs initiales qu'il faut bien fournir par une autre méthode.

- Avec 3 paramètres ou plus la méthode du maximum de vraisemblance ne converge pas toujours.

L'alternative proposée serait en général la méthode des moments classique, mais pourrait-être une méthode des moments aménagée (moments mélangés ou moments de probabilité pondérés) quand des études ont conclu à la supériorité de ces méthodes d'estimation. Dans le logiciel ALED, ce choix a été fait pour la loi des fuites et la loi de Jenkinson.

Des études d'évaluation comparative des divers estimateurs concernant l'une ou l'autre des lois théoriques sont publiées fréquemment et on peut dans un premier temps tenir compte de leurs résultats. Ainsi, pour la loi log-Pearson III, Kishore (A) et Vijay (P.S.) (1989) montrent que les meilleures méthodes sont celles des moments directs et celle des moments mélangés.

6. LES DISTRIBUTIONS EMPIRIQUES ET TESTS D'ADEQUATION DES LOIS AUX DISTRIBUTIONS EMPIRIQUES

6.1. CALCUL DES FREQUENCES EMPIRIQUES.

6.1.1. Rappels théoriques

La fréquence empirique (Plotting position en Anglais) de non dépassement, associée à chaque observation de l'échantillon, découle de son rang quand on classe les observations par valeurs croissantes :

$$x_1 \leq x_2 \leq .... x_{r-1} \leq x_r \leq x_{r+1} \leq .... \leq x_{n-1} \leq x_n$$

La fréquence empirique ou expérimentale de non dépassement correspondant à une valeur x_r est une fonction de r et de n. Il existe plus d'une dizaine d'expressions pour cette fonction.

On peut choisir une expression de manière à obtenir la médiane de la distribution d'échantillonnage des probabilités des valeurs de rang r d'un échantillon de taille n (Michel 1989). L'expression qui en résulte, dite de Chegodayev est :

$$F(x_r) = \frac{r-0.3}{n+0.4}$$

Elle est recommandée par Brunet-Moret (1973) quand les paramètres de la distribution sont connus a priori et elle est utilisée par les services du Ministère Français de l'Agriculture.

On peut choisir une expression qui donne à l'espérance mathématique des valeurs de rang r d'un échantillon de taille n, une probabilité qui est celle de la loi théorique (Cunnane 1978). Les expressions obtenues, qui dépendent donc de la loi théorique, sont souvent compliquées et remplacées par des formules plus simples permettant d'approcher les résultats exacts.

Ces expressions plus simples sont de la forme :

$$F(x_r) = \frac{r-\alpha}{n+1-2\alpha} \text{ avec } 0 \leq \alpha < 0.5$$

On démontre facilement qu'avec la loi uniforme on aboutit à l'expression dite de Weibull :

$$F(x_r) = \frac{r}{n+1}$$

Les variables hydrologiques ayant des distributions généralement bien différentes d'une loi uniforme, cette expression n'est pas recommandée.

Les expressions les mieux adaptées aux variables hydrologiques quand on ne sait pas exactement la loi théorique qui convient sont :

- La formule de Hazen, recommandée par Brunet-Moret (1973) quand les paramètres de la distribution sont estimés à partir de l'échantillon et qui s'écrit :

$$F(x_r) = \frac{r-0.5}{n}$$

- La formule de Cunnane $F(x_r) = \frac{r-0.4}{n+0.2}$ est aussi un bon compromis.

Comme l'indique NERC (1975), les fréquences empiriques sont un guide pour juger de l'adéquation d'une loi théorique et si on ne porte pas une attention spéciale aux valeurs extrêmes, il y a peu de différences entre elles.

6.1.2. Les expressions utilisées par les logiciels

ALED et DIXLOI proposent uniquement l'expression de Hazen. HFA donne le choix entre les quatre expressions citées : Chegodayev, Weibull, Hazen et Cunnane. TROPHEE propose les expressions de Weibull et de Hazen.

6.2. LES TESTS D'ADEQUATION DES LOIS AUX DISTRIBUTIONS EMPIRIQUES

6.2.1. Rappels théoriques

On ne peut guère utiliser en hydrologie les tests dits paramétriques qui font une hypothèse sur la loi théorique, presque toujours la loi normale.

Les tests d'adéquation utilisables quelles que soient les lois utilisées sont donc non paramétriques.

>Le plus connu des tests d'adéquation est le test χ^2 d'ajustement. Bien que peu puissant, il offre l'avantage de fournir une réponse interprétable en terme de probabilité. Le principal reproche qu'on peut lui faire est qu'il nécessite un découpage en classes et que selon la manière de faire ces classes, les résultats peuvent se situer de part et d'autre d'un seuil de signification.

>Brunet-Moret (1978) propose un test dont le principe est le suivant :

- Pour une observation classée x_r de l'échantillon, la loi théorique ajustée permet de calculer une fréquence théorique de non dépassement $F_t(x_r)$, à laquelle on peut faire correspondre une variable normale réduite Vr telle que :

$$F_t(x_r) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{V_r} e^{-t^2/2} dt$$

- On compare ces valeurs V_r à un échantillon idéal de n valeurs U_r rangées par ordre croissant et suivant exactement une loi normale.

- Pour mesurer la distance entre la série des V_r et celle des U_r, Brunet-Moret propose de prendre la somme des carrés des surfaces comprises entre un axe d'abscisses équidistantes :

$$0,\ 2,\ 4,......2(r-1),\ 2r,\ 2(r+1)....2(n-1),\ 2n,\ 2(n+1)$$

et la ligne brisée définie par les points :

$$(2r,\ Z_r) \text{ avec } Z_r = V_r - U_r$$

- Il semble que la probabilité liée à la valeur du test puisse être représentée, dans les tailles d'échantillons comprises entre 8 et 200, par une loi gamma incomplète dont les paramètres s'expriment en fonction de n (Brunet-Moret, 1978).

D'après l'auteur, le point délicat de la construction du Test est la constitution de l'échantillon idéal qui n'a pu être résolue que d'une façon expérimentale.

Les arguments présentés en faveur de ce test et contre le test χ^2, tant par l'auteur que dans la notice d'utilisation de DIXLOI, nous semblent discutables.

Il est reproché au test χ^2, " de renseigner seulement sur la possibilité qu'a la loi choisie avec ses paramètres calculés de représenter la distribution de l'échantillon observé dans la zone de forte densité de probabilité". Autrement dit, de ne pas être influencé par les écarts importants entre les valeurs les plus fortes de l'échantillon et la loi ajustée.

Dans le test de Brunet-Moret au contraire, le poids des valeurs extrêmes est bien supérieur au poids des valeurs centrales.

Ces arguments sont plutôt en faveur du test du χ^2.

Prenons l'exemple d'un échantillon de n = 30 individus représentant la valeur maximale de nombreuses réalisations (pluies journalières par exemple) au cours de n saisons. On montre facilement que l'espérance mathématique de la plus forte valeur de cet échantillon a une période de retour de 54 saisons si on a affaire à une loi de Gumbel. Pour un nombre non négligeable d'échantillons la plus forte valeur aura donc une période de retour bien supérieure à 54 saisons.

On calcule d'ailleurs simplement qu'avec 30 années d'observations, on a une probabilité de $1 - 0{,}99^{30} = 0{,}26$ d'avoir au moins un événement centennal. Sur plusieurs échantillons indépendants d'effectif n = 30, un sur 4 aura au moins un événement centennal.

Il est donc normal d'avoir des points extrêmes qui s'écartent de la loi théorique. Pourquoi vouloir absolument affecter la fréquence de non dépassement 0,984 (correspondant à U_{30} de l'échantillon idéal de Brunet-Moret) à une valeur qui a une probabilité non négligeable d'avoir une fréquence de non dépassement supérieur à 0,99 ? Pour illustrer le fait que de nombreux échantillons de faible effectif peuvent contenir des observations dont la fréquence de non dépassement est très grande, rappelons qu'un orage comme celui qui a frappé la ville de Nîmes le 3 octobre 1988 (plus de 200 mm en quelques heures) a une période de retour ponctuelle de l'ordre de 150 ans, alors qu'il s'observe en moyenne tous les 3 ans sur la région Languedoc Roussillon. De même, sur la région Parisienne on observe en moyenne tous les 8 mois une pluie journalière dont la période de retour ponctuelle est de 10 ans !

>Davis et Stephens (1989) ont présenté une batterie de tests non paramétriques valables dans les conditions suivantes :

- Pour toutes les lois complètement définies (dont les paramètres sont connus a priori),
- Pour la loi normale quand les paramètres sont estimés à partir des observations,
- Pour la loi exponentielle (et donc aussi la loi de Gumbel très voisine) quand les paramètres sont estimés à partir des observations.

Le principe de ces tests est le suivant :

A partir des observations classées par valeurs croissantes,

$$x_1 \leq x_2 \leq \ldots \leq x_i \leq \ldots \leq x_n$$

on calcule :

$Z_i = F(x_i)$ si la loi est complètement définie.

$$Z_i = \Phi\left(\frac{x_i - \bar{x}}{\hat{\sigma}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x_i - \bar{x}}{\hat{\sigma}}} e^{-t^2/2}\, dt$$ si on a affaire à la loi normale.

$Z_i = 1 - e^{-x_i/\bar{x}}$ si on a affaire à la loi exponentielle.

Puis les grandeurs statistiques suivantes sont calculées :

* variable de Kolmogorov-Smirnov = D

$$D = \max(D^+, D^-)$$

$$D^+ = \max\left[(i/n) - Z_i\right] \text{ pour } 1 \leq i \leq n$$

$$D^- = \max\left[Z_i - (i-1)/n\right] \text{ pour } 1 \leq i \leq n$$

* variable de Kuiper = V

$$V = D^+ + D^-$$

* variable de Cramer-Von Mises = W^2

$$W^2 = \sum_{i=1}^{n} \left\{Z_i - \frac{(2i-1)}{2n}\right\}^2 + \frac{1}{12n}$$

* variable de Watson = U^2

$$U^2 = W^2 - n(\bar{Z} - 0{,}5)^2 \text{ avec } \bar{Z} = \sum_{i=1}^{n} \frac{Z_i}{n}$$

* variable d'Anderson-Darling = A^2

$$A^2 = \frac{-\left[\sum_{i=1}^{n}(2i-1)\left\{\mathrm{Ln}\, Z_i + \mathrm{Ln}\,(1-Z_{n+1-i})\right\}\right]}{n} - n$$

Des tables fournissent les valeurs de ces variables pour les 3 cas définis précédemment (loi quelconque définie, loi normale et exponentielle) pour plusieurs risques de première espèce, dans le cas où l'hypothèse Ho est vraie (l'échantillon est tiré de la loi testée).

Pour les lois normale et exponentielle, sous l'hypothèse Ho, il est possible de calculer analytiquement, avec une bonne précision, la probabilité que la variable théorique, W^2, U^2, ou A^2, soit supérieure aux valeurs trouvées sur l'échantillon.

D'après les auteurs, ces tests, qui ne nécessitent pas un découpage en classes, sont plus puissants que le test χ^2.

Michel (1989) recommande aussi le test A^2 d'Anderson-Darling qu'il utilise pour toutes les lois, les paramètres étant estimés sur les observations. Il propose une transformation de A^2 en variable normale réduite, ce qui permet de calculer la fréquence de non dépassement d'une valeur calculée sous l'hypothèse Ho. Le même auteur propose aussi comme test d'ajustement le test du nombre de suites (runs en Anglais). Les points expérimentaux successifs (valeurs observées et fréquences de non dépassement empiriques associées) situés du même côté de la courbe représentant la fonction de répartition théorique, constituent une suite. Si tous les points sont d'un seul côté, il n'y a qu'une suite. Si les points successifs sont alternativement d'un côté de la courbe, puis de l'autre, il y autant de suites que de points. D'après Kendall et Stuart (1943), ce test est sans intérêt quand les paramètres sont estimés sur l'échantillon des observations.

6.2.2. Tests d'ajustement proposés par les logiciels

ALED propose systématiquement le test du χ^2 calculé sur des classes équiprobables, le nombre de classes étant déterminé de manière à ce que leur effectif théorique soit supérieur à 5.

Pour les lois normale, log-normale, exponentielle, de Weibull, de Gumbel et de Jenkinson, il propose aussi les tests W^2, U^2 et A^2. Les résultats sont clairement affichés ainsi que la réponse au seuil de signification 10% retenu : loi acceptée ou rejetée.

DIXLOI, dans un premier affichage avant les sorties graphiques, indique les valeurs prises par les tests de Brunet-Moret et du χ^2, ainsi que le nombre de classes ayant servi à construire ce dernier.
D'après le document qui accompagne le logiciel, le test χ^2 est calculé comme dans ALED.

HFA ne propose aucun test. Dans la documentation il est dit que les tests existants sont rarement suffisamment puissants pour permettre de choisir entre les lois théoriques.
Les tests d'ajustement réalisés par TROPHEE sont :

- le test du χ^2
- le test d'Anderson-Darling mis en oeuvre selon la méthode de Michel (1989)
- le test de Kolmogorov-Smirnov,
  pour trois valeurs de risque de 1ère espèce : 1%, 5%, 10%.

Le nombre de classes sur lesquelles est effectué le test du χ^2 , est calculé par une méthode différente de celle de DIXLOI et ALED. Ce test n'est utilisé qu'à partir d'un nombre de valeurs supérieur à 20.

6.2.3. Propositions

Bien que peu puissant, le test du χ^2 doit être maintenu, mais il faut afficher les résultats de manière claire comme le fait ALED où il manque encore l'affichage du nombre de classes et la possibilité de choisir le seuil de rejet au début du travail.

Des trois tests W^2, U^2 et A^2 qui donnent des résultats probablement redondants, on pourrait ne retenir que le test A^2 d'Anderson-Darling au moins pour les mêmes lois qu'ALED. Il faudrait voir dans les publications spécialisées si des méthodes n'ont pas été développées pour utiliser ce test, quelle que soit la loi théorique, quand les paramètres sont estimés sur les observations. On pourrait tester la transformation proposée par Michel (1989) pour calculer la fonction de répartition de A^2 sous l'hypothèse Ho dans ces conditions (par simulation des différentes lois). A priori il est étonnant qu'une transformation qui ne fait pas intervenir le nombre de paramètres des lois (comme le nombre de degrés de libertés de la distribution χ^2) puisse convenir dans tous les cas.
Enfin comme le souligne clairement le document d'accompagnement du logiciel TROPHEE, les tests d'ajustement ne sauraient à eux seuls permettre de trancher entre les

ajustements réalisés par plusieurs lois. Tout au plus, servent-ils à donner des indications supplémentaires sur la qualité des ajustements.

Pour juger de la qualité d'un ajustement, il ne faut donc pas oublier de regarder les graphiques où, en même temps que la fonction de répartition théorique de la loi ajustée, sont portés les points correspondants aux couples : valeur observée - fréquence de non dépassement empirique associée.

7. LES REPRESENTATIONS GRAPHIQUES

Elles sont de deux sortes :
- L'histogramme des valeurs observées
- La fonction de répartition de la loi théorique, représentée par une courbe continue en même temps que les points correspondant aux couples valeur observée - fréquence empirique de non dépassement, et associée éventuellement aux courbes continues correspondant à un intervalle de confiance pour les quantiles.

ALED propose un histogramme des valeurs observées (cf. § 3.4).

Il propose toujours une représentation arithmétique de la fonction de répartition de la loi théorique et des points expérimentaux. Sur cette répartition arithmétique, la variable est en abscisse et la probabilité (ou fréquence) de non dépassement en ordonnée.

Pour les lois normale, log normale, de Gumbel et de Jenkinson, une représentation avec des graduations fonctionnelles permet la linéarisation des lois à 2 paramètres. Dans ce cas, la probabilité est en abscisse et la variable en ordonnée.

Les intervalles de confiance (cf.§ 8) quand ils sont calculés, ne sont jamais représentés graphiquement.

Un titre (60 caractères au maximum) est répété systématiquement sur chaque graphique. Pour les graphiques concernant la fonction de répartition figurent le nom de la loi et la méthode d'ajustement. Les unités de la variable ne sont pas clairement indiquées. Il n'y a pas possibilité de modifier les informations, ni d'ajouter des légendes.

Il n'est pas possible de représenter les fonctions de répartition de plusieurs lois sur le même graphique.
Il existe deux versions du logiciel : une pour la carte graphique Hercule, une autre pour les cartes graphiques CGA, EGA et VGA. Dans cette dernière version, les possibilités graphiques des cartes EGA et VGA ne sont pas exploitées car les graphiques ont toujours la définition la plus faible liée à la carte CGA.
Après avoir chargé deux programmes DOS fournis avec le logiciel, la touche PRINT SCREEN permet d'imprimer les graphiques sur presque toutes les imprimantes matricielles courantes.

Dans **DIXLOI**, il n'y a pas d'histogramme des valeurs observées.
Indépendamment de la bizarrerie qui oblige à quitter le menu pour voir apparaître les fonctions de répartition à l'écran, l'environnement graphique est assez riche. Le titre de l'échantillon apparaît. Les probabilités de non dépassement sont portées en abscisse et les valeurs de la variable en ordonnée. On peut ou non faire apparaître un quadrillage pour faciliter la lecture, et choisir une graduation fonctionnelle parmi trois pour les probabilités (arithmétique, gaussienne et Gumbelienne). Remarquons toutefois que la graduation gaussienne des probabilités ne suffit pas pour rendre linéaire la loi log-normale.

On peut ajouter du texte au titre et en principe ajouter une légende aux axes et faire apparaître les valeurs des paramètres des lois ajustées.
Un cartouche que l'on peut déplacer en n'importe quel endroit de l'écran, ou faire disparaître, indique la loi ajustée, la méthode d'ajustement et rappelle les principales caractéristiques de l'échantillon (moyenne, écart-type, nombre de points...).
Le tracé des intervalles de confiance n'est pas opérationnel.
Trois courbes au plus peuvent figurer sur le même graphique. Une copie d'écran n'est pas possible.
Une option du menu général prévoit le tracé des graphiques sur table traçante.

HFA permet certaines représentations graphiques liées à la mise en oeuvre des tests de vérification de la qualité des échantillons (test d'homogénéité de Mann-Whitney, et test de Grubbs et Beck de détection des outliers (cf. § 3.3.2.)). Les graphiques des fonctions de répartition ont des graduations fonctionnelles qui linéarisent les lois à deux paramètres. Les deux graduations de probabilité proposées sont :

- celle relative à la loi normale
- celle relative à la loi gamma.

Des transformations simples permettent, à partir de ces deux seules graduations de probabilité, la linéarisation de toutes les lois à deux paramètres.

La probabilité cumulée au non-dépassement est portée en abscisse, la variable ou son logarithme en ordonnée. La nature de la graduation utilisée est précisée. Quatre courbes au maximum sont portées sur le même graphique. Toute courbe isolée peut être représentée avec son intervalle de confiance pour un seuil de confiance au choix de 50, 80 ou 95 %.

La clarté des graphiques est remarquable. Un quadrillage facilite la lisibilité des courbes. Légende et titre y figurent très clairement.

Tous les graphiques sont construits à partir de la carte graphique incorporée dans le micro-ordinateur.

L'impression sur imprimante (IBM, EPSON, HP compatible) ou table traçante (Plotter HP et compatible) est opérationnelle.

TROPHEE propose les représentations graphiques des fonctions de répartition empirique et théorique.
En abscisse deux échelles sont représentées. L'une porte les fréquences cumulées au non dépassement suivant une graduation fonctionnelle, l'autre les périodes de retour correspondantes.

La variable étudiée est représentée sur l'axe des ordonnées. L'échelle est arithmétique. Les titres du graphique sont modifiables. Le nom des lois ajustées et les méthodes d'ajustement utilisées sont notées sur le graphique, de même que la période que recouvrent les observations. Trois courbes au maximum sont représentées sur le même graphique. Si une seule courbe est représentée, les valeurs des paramètres figurent sur le graphique.

Le programme calcule automatiquement l'échelle de l'axe des ordonnées de manière à ce que le tracé occupe le maximum de place. Des options permettent de modifier le tracé proposé par défaut :

- une échelle manuelle pour l'axe des ordonnées et son intitulé peut être définie,

- le tracé des points observés peut être désactivé,

- si une loi est représentée par graphique, il y a possibilité de tracer l'intervalle de confiance à 90%,

- des tableaux des quantiles principaux en année humide et/ou en année sèche (périodes de retour 5, 10, 20, 50, 100 ans), peuvent figurer à droite du graphique.

La dernière option concerne la représentation de l'estimation de la période de retour d'une valeur donnée.

Les graphiques peuvent être imprimés ou tracés sur table traçante. Les périphériques de sortie utilisés sont déclarés dans un fichier.

Remarque : nous n'avons pas étudié les différents modes de gestion éventuels de fichiers graphiques retenus par les logiciels qui relèvent de considérations d'ordre strictement informatique.

Propositions

Les graphiques concernant l'échantillon des observations sont intéressants de manière optionnelle. On devrait pouvoir sortir :

- la chronologie des valeurs observées quand celles-ci sont associées à une date,

- la distribution de fréquence en fonction du mois quand les dates sont indiquées,

- l'histogramme dans tous les cas.

Les graphiques des fonctions de répartition devraient avoir les probabilités de non-dépassement en abscisse et les valeurs de la variable en ordonnée. En outre :

- Le choix des graduations fonctionnelles devrait se faire automatiquement de manière à linéariser les lois à 1 paramètre (exponentielle et géométrique linéarisables en 1-F (x)) et la plupart des lois à 2 paramètres (la méthode est à mettre au point pour la loi des fuites). La loi binomiale négative tronquée serait portée dans les mêmes repères que la loi géométrique, et la loi de Weibull dans les mêmes repères que la loi exponentielle.

Les lois à plus de 2 paramètres seraient portées dans le système qui linéarise la loi à 2 paramètres correspondante (de manière à juger visuellement de l'intérêt d'un 3ème paramètre).

- Quand plusieurs lois sont portées sur le même graphique (possibilité qui doit être prévue), une seule graduation fonctionnelle sera retenue, par exemple celle de la loi qui a le meilleur résultat aux tests d'ajustement.

- On devra pouvoir faire apparaître ou non un quadrillage pour faciliter la lecture, et tracer ou non les intervalles de confiance.

- Automatiquement et systématiquement devraient apparaître un titre minimum (pour identifier le graphique), le nom de la loi ou des lois et des méthodes d'ajustement, mais on devrait pouvoir les déplacer sur le graphique, rajouter du texte au titre, préciser les légendes des axes, donner éventuellement les caractéristiques de l'échantillon ou les paramètres de la loi.

Enfin, il devrait être possible de manière simple de faire sortir les graphiques sur une imprimante matricielle, à jet d'encre ou laser et sur table traçante.

8. LES INTERVALLES DE CONFIANCE DES QUANTILES

Un quantile Xp est une valeur de la variable aléatoire X telle que :

$$F\ (xp) = \text{Prob}\ (X \leq xp) = p$$

Pour une valeur de p fixée, la forme analytique de la fonction de répartition F (x) et de sa dérivée par rapport à x, la fonction densité de probabilité f (x), ainsi que les valeurs des paramètres α, β, γ qu'elles font intervenir, permettent le calcul de xp.xp est tel que :

$$\int_{-\infty}^{xp} f(x,\ \alpha,\ \beta,\ \gamma)dx = p$$

Les valeurs de α, β, γ sont estimées à partir d'un échantillon d'observations. Si on pouvait disposer de plusieurs échantillons tirés de la même population, on verrait que chacun d'entre eux donne un jeu de paramètres différents, donc une valeur différente de xp qui n'est en définitive qu'un estimateur de la vraie valeur inconnue du quantile.

On tient compte des fluctuations d'échantillonnage en associant à p non plus une valeur ponctuelle xp, mais un intervalle :]xp-d1 ; xp + d2[qui a une probabilité 1-α de recouvrir la vraie valeur du quantile : c'est l'intervalle de confiance.
La probabilité 1- α est appelée seuil ou degré de confiance et on lui donne le plus souvent les valeurs 90 ou 95%.

Pour calculer l'intervalle de confiance du quantile, il faut connaître sa distribution d'échantillonnage. Cramer (1946) donne une expression de la densité de probabilité de la distribution d'échantillonnage de Xp, qui fait intervenir la loi de X. Cependant :

- L'intégrale de cette expression ne peut être obtenue que par des méthodes numériques approximatives et compliquées.

- Le caractère rigoureux de la démarche ne se justifie pas en hydrologie parce que le choix d'une loi ne résulte pas d'une théorie, mais d'une démarche expérimentale.

Une étude par simulation a permis à Kite (1975) de montrer, en ce qui concerne les lois de probabilité habituellement utilisées en hydrologie, que la distribution d'échantillonnage des quantiles n'est pas significativement différente d'une loi normale.

L'estimation de la moyenne $\mu(Xp)$ et de l'écart-type $\sigma(Xp)$ de la variable aléatoire Xp permet donc de calculer la valeur de la variable centrée réduite U :

$$U = \frac{(Xp-\mu(Xp))}{\sigma(Xp)}$$

dont l'intervalle de confiance est :

$$\text{Prob } (U_{\alpha/2}<U<U_{1-\alpha/2}) = 1 - \alpha$$

$U_{\alpha/2}$ et $U_{1-\alpha/2}$ sont des quantiles de la loi normale réduite. A cause de la symétrie de la loi normale, on a : $U_{\alpha/2} = -U_{1-\alpha/2}$

En remplaçant U par son expression, on obtient les limites de confiance de Xp qui sont :

$$\mu\ (Xp) \pm U_{1-\alpha/2}\ \sigma(Xp)$$

La seule estimation possible de μ (Xp) est la valeur xp qu'on calcule. L'expression des limites de l'intervalle de confiance est en définitive :

$$xp \pm U_{1-\alpha/2}\ \sigma(Xp)$$

Estimation de la variance du quantile

La technique d'estimation dépend des méthodes qui ont été utilisées pour estimer les paramètres de la loi théorique.

Si c'est la méthode des moments, alors le quantile xp peut s'exprimer en définitive comme une fonction de ces moments :

$$xp = g\ (\hat{\mu}'1, \hat{\mu}2, \hat{\mu}3, p)$$

$\hat{\mu}'1$: estimateur de la moyenne de la population
$\hat{\mu}2$: estimateur de la variance de la population
$\hat{\mu}3$: estimateur du moment centré d'ordre 3 de la population

A partir d'un développement en série de Taylor autour des valeurs moyennes des moments, en ne retenant que les premiers termes du développement, on obtient une relation linéaire qui permet un calcul facile mais approximatif de la variance de xp. Le calcul fait intervenir :

- Les dérivées de la fonction g par rapport aux différents moments. Ces dérivées sont spécifiques d'une loi de probabilité théorique donnée.

- Les variances et covariances des estimateurs des différents moments, dont on trouve les expressions dans Kendall et Stuart (1943).

En ce qui concerne la méthode du maximum de vraisemblance, la technique de calcul de la variance du quantile est très voisine. Le quantile s'exprime cette fois en fonction des paramètres de la loi :

$$xp = g\ (\hat{\alpha}, \hat{\beta}, \hat{\gamma}, p)$$

Un développement en série de Taylor autour des valeurs moyennes de ces paramètres dont on ne retient que les premiers termes fournit là encore un moyen commode mais approximatif pour calculer la variance de xp.

Comme précédemment, le calcul fait intervenir les dérivées de la fonction g par rapport aux différents paramètres, dérivées qui dépendent de la loi de probabilité.

Les variances et covariances asymptotiques des estimateurs des paramètres sont fournies par la matrice de dispersion dont on trouve la justification théorique dans Kendall et Stuart (1983). Pour des exemples d'applications pratiques on pourra consulter Masson (1983).

On remarque :

- que le calcul de l'intervalle de confiance est approximatif en ce qui concerne l'estimation de la variance du quantile et de sa distribution d'échantillonnage supposée normale, même quand l'effectif de l'échantillon est petit.

- que le calcul de l'intervalle de confiance suppose que la loi théorique convient, ce qui est choquant à première vue. A la réflexion, on procède toujours de la sorte : tous les tests supposent l'hypothèse Ho vraie et rares sont les exemples où on se préoccupe du risque de deuxième espèce (risque de retenir Ho alors que c'est une autre hypothèse H, dite alternative, qui convient !).

Les logiciels DIXLOI, HFA et TROPHEE calculent les intervalles de confiance selon les principes indiqués ci-dessus. Le logiciel ALED ne calcule l'intervalle de confiance d'un quantile que pour la loi de Weibull.

Proposition

La méthode décrite ci-dessus est celle qui est utilisée le plus généralement pour calculer les intervalles de confiance et elle convient tout à fait.
Le Ministère de l'Agriculture propose une autre méthode que nous n'avons pas retenue faute d'en avoir compris les bases théoriques.

L'opérateur devrait pouvoir choisir le degré de confiance $1 - \alpha$ et décider de visualiser ces intervalles de confiance soit sous forme graphique, soit sous forme de tableaux, pour un nombre fini de valeurs de périodes de retour données.

9. FICHIERS RESULTATS

Les résultats des ajustements réalisés sont parfois sauvegardés dans des fichiers ASCII.

ALED ne prévoit pas de stockage fichier des résultats. Ces derniers sont seulement visualisés à l'écran. Il s'agit :

- des valeurs des paramètres des lois ajustées, la méthode d'ajustement étant précisée, et de la table des quantiles pour les probabilités de non dépassement de 0.01 à 0.99 par pas de 0.01. Une option permet par ailleurs d'obtenir la valeur de la variable correspondant à n'importe quelle probabilité de non dépassement rentrée au clavier.

DIXLOI : les résultats sont obligatoirement stockés dans un fichier dont le contenu peut ensuite être imprimé. Deux types de fichiers sont proposés, l'un dit condensé, l'autre dit complet contenant toutes les informations du premier et des compléments particuliers.

Les informations communes aux deux fichiers sont :

- les caractéristiques de l'échantillon,

- pour chaque loi ajustée, les valeurs des paramètres estimés, les caractéristiques de la population, les fréquences théoriques de non dépassement associées aux valeurs observées, les résultats des tests d'adéquation des ajustements réalisés.

Le fichier complet contient en outre :

- le tableau des données observées par ordre d'apparition,
- la fonction de répartition empirique,
- pour chaque loi ajustée, les quantiles et intervalles de confiance relatifs à des valeurs de fréquence de non dépassement choisies au moment de la définition des conditions de l'ajustement. Les périodes de retour correspondantes sont données également.

HFA

- Les valeurs des paramètres des lois ajustées, la méthode d'ajustement étant précisée sont visualisées à l'écran.
  Des touches de fonction permettent alors d'imprimer les résultats ou de les sauvegarder dans un fichier.

- Celui-ci contient,
  - la fonction de répartition empirique avec le rappel de la formule de plotting position utilisée,
  - les caractéristiques essentielles de l'échantillon sur les valeurs brutes et sur les logarithmes décimaux des valeurs brutes,

- pour chaque loi ajustée,
  - les paramètres estimés et la méthode d'ajustement utilisée,
  - les caractéristiques théoriques de la population,
  - les quantiles correspondants à des valeurs de fréquences de non-dépassement (dont 0.001 - 0.005 - 0.01 - 0.99 - 0.995 - 0.999), les écarts-types de ces quantiles, et les intervalles de confiance à 50%, 80% et 95%.

TROPHEE

Les résultats des ajustements ne figurent à l'écran que sur les graphiques.
Toutefois une option permet d'obtenir une impression de résultats détaillés qui sont par ailleurs stockés dans un fichier temporaire. Ce fichier est en effet effacé lors de chaque nouvelle exécution du module d'ajustements statistiques.

Il contient :

- la fonction de répartition empirique avec le rappel de la formule de plotting position utilisée,
- les caractéristiques essentielles de l'échantillon
- pour chaque loi ajustée,
  - le rappel de l'expression analytique de la fonction densité de probabilité,
  - les valeurs théoriques correspondant aux fréquences empiriques des valeurs observées,
  - quelques quantiles caractéristiques et leurs intervalles de confiance à 90%,
  - les résultats des tests d'ajustement.

Une option permet d'obtenir à l'écran la valeur de la variable correspondant à n'importe quelle période de retour rentrée au clavier et réciproquement .

<u>Proposition</u>

Les résultats à retenir sont :

- la fonction de répartition empirique avec le rappel de la formule de plotting position utilisée,
- les caractéristiques essentielles de l'échantillon,
- les conclusions des tests de vérification de la qualité des ajustements.

- pour chaque loi ajustée,

  - le rappel de l'expression analytique de la fonction densité de probabilité et la définition des paramètres,
  - la méthode d'ajustement utilisée,
  - les paramètres estimés,
  - le calcul des quantiles et des intervalles de confiance correspondants pour des probabilités de non dépassement préalablement définies,
  - les résultats des tests d'adéquation des ajustements.

Une option doit permettre d'obtenir à l'écran la valeur de la variable correspondant à n'importe quelle probabilité de non dépassement, et réciproquement.

10. DOCUMENTS D'ACCOMPAGNEMENT DES LOGICIELS

Ces documents d'accompagnement sont indispensables.

Dans le cas des logiciels étudiés :

ALED ne possède pas de documentation spécifique. Un message écran précise les objectifs et les limites du logiciel.

Toutefois, il faut préciser que ce logiciel est destiné à des élèves-ingénieurs qui en disposent pour faire des études hydrologiques dans le cadre de projets d'étude, et que les documents pédagogiques suivants sont à leur disposition :

- La loi log-normale (32 pages) (Masson, 1985)
- La loi de Gumbel (40 pages) (Masson, 1983)
- La loi gamma incomplète (36 pages) (Masson, 1982)
- La loi des fuites (21 pages) (Masson, s.d.)
- La loi de Weibull (11 pages) (Masson, s.d.)
- Le calcul des intervalles de confiance (41 pages) (Masson, 1983)

Dans ces brochures sont exposées les méthodes, et détaillées les différentes techniques de calcul.

DIXLOI est livré avec une notice d'utilisation de 56 pages (Lebel et Boyer, 1989). On y trouve des généralités statistiques, un mode d'emploi du logiciel qui laisse parfois l'utilisateur dans l'embarras, mais rien de précis sur les techniques de calcul utilisées pour estimer les paramètres de telle ou telle loi. Des références indiquées, seule Kite (1976) permet d'identifier des méthodes et techniques de calcul utilisées, mais ne couvre pas l'ensemble des lois traitées.

HFA est livré avec un livre qui traite dans le détail des lois gamma à la fois sous l'aspect théorique et pratique. On y trouve non seulement les formulations analytiques, les différentes méthodes d'estimation des paramètres, les liaisons avec les autres lois mais aussi, à partir d'un exemple, les résultats détaillés des différentes méthodes : il est possible de vérifier pas à pas les résultats donnés par le logiciel.

Par ailleurs, un manuel d'utilisation complet accompagne le logiciel.

TROPHEE est accompagné d'une notice d'utilisation. Y figure un organigramme détaillé décrivant l'enchaînement des différentes tâches du module de traitements statistiques. L'utilisation des touches de fonctions et des différentes options est largement explicitée. Du point du vue statistique par contre aucune précision n'est apportée sur les algorithmes de calcul des paramètres estimés, des quantiles, des intervalles de confiance et des périodes de retour.
Le recours au code FORTRAN est nécessaire pour recueillir ces informations.

Seuls les tests d'ajustement méritent des compléments méthodologiques en annexe. Les références sont pratiquement inexistantes.

Propositions

Les logiciels d'ajustement devraient être accompagnés de 2 types de documents.

1 - Un document fourni systématiquement avec le logiciel et qui serait un mode d'emploi détaillé. Il ne rentrerait pas dans le détail des méthodes mais expliquerait comment accéder aux différentes fonctions du logiciel. Il serait illustré par un ou plusieurs échantillons tests (fournis avec le logiciel) avec des exemples d'utilisation et de résultats. On pourrait s'inspirer de logiciels travaillant dans d'autres domaines dont on aurait apprécier la documentation particulièrement bien faite.

2 - Un document beaucoup plus important et fourni seulement en option où figureraient :

   - Des rappels statistiques sommaires, mais clairs et rigoureux, concernant les grandeurs statistiques calculées, les méthodes d'estimation de paramètres et de calcul des intervalles de confiance.

   Pour chaque loi ajustée :

   - Un rappel de la formulation et des paramètres calculés par le logiciel. Si s est un paramètre, $a = \frac{1}{s}$ est aussi un paramètre et il n'est pas inutile de préciser lequel des deux utilise le logiciel.

- Un rappel des expressions mathématiques utilisées pour estimer les paramètres avec les algorithmes détaillés du calcul numérique quand il n 'y a pas de solution explicite.

- Un rappel des expressions mathématiques utilisées pour calculer les quantiles et les intervalles de confiance.

- Un guide des principales méthodes qui permettent de passer des résultats d'un ajustement à l'estimation du risque.

Ce document nous semble indispensable parce qu'il permet de savoir ce que fait précisément le logiciel sans décrypter le programme source. Il pourrait se présenter sous forme d'un ensemble de fascicules détachables. En effet :

* Les méthodes sont susceptibles d'évoluer et une méthode de calcul plus performante pourra remplacer la méthode de calcul retenue lors de la fabrication du logiciel.

* De nouvelles méthodes d'ajustement des paramètres, de nouveaux tests d'ajustement sont susceptibles d'apparaître.

* De nouvelles lois peuvent être ajoutées au logiciel.

* L'estimation du risque à partir des résultats d'un ajustement dépend du problème posé et de l'échantillon des données. Il n'est guère possible de faire un inventaire exhaustif de tous les cas possibles au départ. Le document pourrait être complété au fur et à mesure des problèmes posés par les hydrologues.

11. UN LOGICIEL D'AJUSTEMENT DE LOIS STATISTIQUES ENTRE LA BOITE NOIRE ET LE SYSTEME EXPERT

11.1. LES LOIS ET LEUR AJUSTEMENT

On peut penser que l'utilisateur averti des logiciels qui viennent d'être étudiés choisit parmi les lois proposées celles qui sont susceptibles de convenir à l'échantillon de données dont il dispose.

Toutefois, il n'est pas rare, nous avons pu le constater, que le projeteur d'une étude hydrologique suivant l'exemple de la notice d'utilisation (DIXLOI) ajuste toutes les lois proposées.
Cette manière de procéder est assez choquante dans la mesure où le logiciel est utilisé comme un appareil automatique dans lequel on place à une extrémité le matériel d'observation pour recueillir à l'autre extrémité la solution supposée exacte du problème.

Par ailleurs, un certain nombre de lois de probabilité à deux paramètres ne tolèrent que des valeurs strictement positives. Il en est ainsi des lois exponentielle, de Weibull, et gamma, sans parler du cas plus trivial des transformations logarithmiques. D'autres lois comme la loi des fuites sont bien adaptées à la présence de valeurs nulles.

Les lois à trois paramètres échappent à ce problème, mais en présence de valeurs nulles, le paramètre de position si on le laisse s'ajuster peut prendre une valeur négative, ce qui permet de calculer une probabilité pour des valeurs de la variable négatives et supérieures au paramètre de position, même si cela n'a pas de sens. Cette constatation n'a rien d'étonnant : le paramètre de position ajusté assure l'égalité des moments ou le maximum de vraisemblance mais ignore les contraintes physiques !

Pour ajuster des lois qui font intervenir une transformation logarithmique sur des données comportant quelques valeurs nulles, certains projeteurs transforment ces valeurs nulles en valeurs presque nulles, généralement 0.1 - 0.01 ou 0.001.
On peut remarquer que 1.10^{-30} est plus proche de zéro que 1.10^{-1}. Mais selon que l'on retient l'une ou l'autre valeur, les résultats de l'ajustement diffèrent sensiblement et ce genre de pratique est à éviter.

Il résulte de ces remarques une réflexion plus générale sur le paramètre de position des lois à 3 paramètres qui, en théorie, correspond à la borne inférieure de la variable mais qui, dans les faits, n'est le plus souvent qu'un paramètre d'ajustement.

Le logiciel DIXLOI laisse à l'utilisateur le choix de fixer ce paramètre de position (par exemple si on juge que la borne inférieure est physiquement zéro, ou si on a sélectionné des observations supérieures à un seuil), de le borner ou de l'obtenir par ajustement. ALED demande le seuil au-dessus duquel on considère la variable quand on ajuste une loi de Weibull (ce seuil peut être zéro), et affecte cette valeur au paramètre de position de la loi.

Proposition

On pourrait concevoir un logiciel à 2 options :

- une option libre, pour hydrologue averti, qui sait où il va et qui connait les voies à emprunter pour y arriver .

- Une option guidée pour le projeteur qui veut obtenir rapidement un résultat sans trop se poser de questions.

Ce guidage ne serait pas contraignant mais indicatif. Sans nous livrer à une étude détaillée du problème, nous pouvons indiquer quelques pistes :

* grâce à l'erreur standard de γ_1 et γ_2 sur les variables brutes ou transformées en logarithmes, on peut admettre ou non les lois normales ou log-normales à 2 paramètres.

* En présence de valeurs nulles dans l'échantillon, on déconseillerait (ne faudrait-il pas interdire) à l'utilisateur d'ajuster des lois à 2 paramètres qui ne les tolèrent pas.

* En ce qui concerne le paramètre de position il semble intéressant de disposer des deux possibilités suivantes :

  - fixer le paramètre de position,
  - ajuster le paramètre de position.

  Si l'une des options ne donnait pas satisfaction, il faudrait alors recommencer l'ajustement avec l'autre.

* La loi de Gumbel suppose un coefficient de dissymétrie de 1.139. Par simulation (Rossi, Fiorentino et Versace, 1984) on devrait trouver des limites, fonction de la taille de l'échantillon, au-delà desquelles la loi de Gumbel ne convient pas.

* Certaines lois (géométrique, binomiale négative tronquée) concernent des variables discrètes supérieures ou égales à 1. Leur utilisation serait déconseillée sur des variables continues.

* Les échantillons présentant un histogramme en i sont susceptibles de voir leur distribution bien ajustée par une loi exponentielle ou de Weibull. On orienterait vers ces lois les échantillons présentant de tels histogrammes.

Il n'en demeure pas moins qu'avec un histogramme en cloche très dissymétrique, on pourra hésiter entre plusieurs lois à 3 paramètres : log-normale, gamma et Jenkinson par exemple. C'est à ce niveau que les habitudes consacrées par l'usage, l'expérience et le jugement de l'utilisateur doivent jouer.

11.2. ESTIMATION DU RISQUE

Le résultat d'un ajustement ne donne pas toujours une estimation directe du risque.

Pour illustrer l'écart qui sépare le résultat d'un ajustement de l'estimation du risque, prenons un exemple simple et fréquent : on s'intéresse aux pointes de crue d'une rivière à un endroit précis de son cours et on veut associer à différentes valeurs de ce débit de pointe une période de retour exprimée en années.

a) Si on retient le débit maximum observé chaque année, la probabilité de non dépassement F (x) donnée par la loi ajustée est une probabilité annuelle et on obtient directement la période de retour T associée à une valeur x du débit en faisant :

$$T = \frac{1}{1-F(x)}$$

b) Si on retient tous les débits maximaux des crues indépendantes dépassant un seuil donné, la valeur F (x) fournie par la loi ajustée n'est pas la probabilité annuelle de non dépassement. Il faut la corriger du nombre moyen annuel de dépassement du seuil :

$$T = \frac{1}{\lambda[1-F(x)]}$$

avec

$\lambda = \frac{n}{N}$ nombre moyen annuel de dépassements

n = nombre de dépassements du seuil
N = nombre d'années d'observations.

c) Si on retient les k plus forts débits indépendants observés chaque année, la valeur de F(x) donnée par la loi ajustée n'est pas une probabilité annuelle de non dépassement ; c'est $[F(x)]^k$ qui est approximativement une probabilité annuelle de non dépassement et on a :

$$T \approx \frac{1}{1-[F(x)]^k}$$

d) Si on se trouve sur une rivière des Alpes du Sud on pourra distinguer deux saisons : le printemps avec des crues de fonte des neiges fortes surtout en volume, et l'automne avec des crues beaucoup plus pointues dues à des orages.

On obtiendra la loi de probabilité des débits maximaux d'une saison par une des 3 méthodes ci-dessus (a, b, ou c). Ces probabilités de non dépassement par saison F_1 (x) et F_2 (x) étant déterminée, on obtiendra la probabilité de non dépassement annuelle en faisant le produit des deux probabilités précédentes et la période de retour T associée à une valeur de débit x sera :

$$T = \frac{1}{1-F_1(x)*F_2(x)}$$

Cet exemple ne concerne qu'une variable simple : le débit de pointe. Si l'on s'intéresse à l'ensemble d'une crue définie par son hydrogramme, ou l'ensemble d'une averse définie par son hyétogramme, le problème est autrement complexe.
Les logiciels DIXLOI et TROPHEE donnent une estimation du risque dans les cas a et b.

Proposition

Le logiciel pourrait traiter des cas simples sur la base de ceux qui viennent d'être évoqués par un échange de questions-réponses avec l'utilisateur pour définir précisément la nature du problème à résoudre et de la variable étudiée. Devant un cas de figure non répertorié, le logiciel donnerait un message précisant que l'estimation du risque nécessite une étude particulière.

Ce type de développement fait peut-être appel à une approche type système-expert.

12. ETUDES PONCTUELLES POUR L'EXTENSION ET LE DEVELOPPEMENT DU LOGICIEL

Un logiciel n'est pas un produit figé. On lui donne même un numéro qui augmente avec le degré de perfectionnement (version 6.0 de TURBO PASCAL par exemple).

De nouvelles méthodes de calcul, de nouvelles lois sont susceptibles d'être ajoutées au logiciel d'ajustement. Ces ajouts ne doivent pas être faits à l'impulsion, mais après avoir été testés par des études ponctuelles. Beaucoup d'entre elles pourraient être faites par des stagiaires bien encadrés.

Par ailleurs des problèmes restent en suspens et méritent d'être abordés.
Parmi les études que la rédaction de ce rapport nous a amenés à envisager, nous avons noté :

Echantillons censurés et lois tronquées.

Pour décrire la distribution d'une variable hydrologique, il n'est pas rare d'utiliser une loi théorique définie sur un domaine plus large que celui des observations.

Un exemple est donné par la loi normale définie sur le domaine $] -\infty , +\infty [$ que l'on ajuste sur des totaux annuels de précipitations qui ont une probabilité nulle en 0 et le plus souvent pour des valeurs supérieures à 0.

Doit-on en toute rigueur tronquer la loi de distribution au seuil x_0 en-deçà duquel aucune valeur n'a été observée sur l'échantillon ? Cette façon de procéder a-t-elle une influence notable sur l'estimation des paramètres, des probabilités et des risques, par rapport à un ajustement classique ?

Doit-on ajuster des lois à 3 paramètres en fixant le paramètre de position au seuil x_0 ?

Par ailleurs, on dispose parfois d'échantillons où les faibles ou fortes valeurs ne sont pas mesurées quantitativement. Il s'agit par exemple de précipitations inférieures au seuil de détection de 0.1 mm des pluviomètres, ou des chutes de pluie supérieures à la capacité du pluviomètre. Un problème intéressant est de savoir comment estimer au mieux la loi de probabilité d'une variable dont on possède un tel échantillon dit censuré (Kendall et Stuart, 1943) : par exemple n années d'observations de débits et des informations concernant les crues historiques au cours des 150 dernières années.

Si on s'intéresse aux probabilités d'occurence de valeurs du domaine censuré les techniques d'ajustement doivent être adaptées à cette situation. En particulier, la fonction de vraisemblance à maximiser s'exprime de manière à tenir compte des valeurs censurées en nombre connu.

Par contre, si seules les valeurs situées loin du point de censure font véritablement l'objet de l'étude statistique, est-il important de connaître la proportion de valeurs égales ou inférieures au seuil de censure ? En d'autres termes la connaissance de cette proportion permet-elle de préciser la distribution des valeurs "éloignées" de x_0 ? Dans ce cas les résultats obtenus à partir d'un ajustement, par une méthode habituelle, d'une loi théorique à toutes les valeurs dépassant un seuil donné, ne sont-ils pas tout à fait valables ?

Ainsi est-il important de connaître avec précision les proportions de pluies égales ou inférieures à un seuil de l'ordre de quelques millimètres pour déterminer la distribution des pluies les plus fortes ?

Inversement on utilise parfois une loi théorique dont le domaine de définition est plus petit que celui de la variable : par exemple une loi log-normale à 2 paramètres sur une série de pluies journalières comprenant des valeurs nulles. On peut envisager de traiter le problème en enlevant de l'échantillon les observations non admises par la loi puis en procédant à un ajustement classique sur les données restantes. Les probabilités correspondantes sont ensuite corrigées de la proportion calculée sur l'échantillon des valeurs non retenues afin de définir une fonction de répartition comprise entre 0 et 1 sur l'ensemble du domaine des observations.

Mais on peut aussi choisir une autre loi tolérant l'ensemble des valeurs et qui donnerait donc une estimation théorique de la probabilité des valeurs non prises en compte précédemment (loi des fuites par exemple).

Plus généralement c'est de l'étude des échantillons censurés et de l'application des lois tronquées (terminologie définie par Kendall et Stuart, 1943) dont il est question.

Il y aurait peut-être toute une reflexion à mener sur ce thème autour duquel règne une certaine confusion (Lubès, 1992).

Comment traiter les valeurs extrêmes inférieures (étiages)

Comment faire pour adapter aux valeurs inférieures les lois habituellement utilisées pour les valeurs extrêmes supérieures ? Y-a-t-il des lois spécifiques aux valeurs extrêmes inférieures ?

Loi des valeurs extrêmes à 2 composantes

Recensement des cas où elle est susceptible de s'appliquer : régions avec des pluies de nature différente (cyclones tropicaux), fleuves avec des crues débordantes ou non (Niger ?) etc...

Quelles valeurs initiales prendre pour estimer les paramètres par la méthode du maximum de vraisemblance. Quel algorithme assure la convergence vers la solution ? quelles sont les conditions pour qu'il y ait convergence ?

Que penser de quelques lois utilisées par d'autres organismes

La loi somme de 2 exponentielles utilisée par EDF pour étudier la distribution des hauteurs de pluie sur des durées comprises entre 2 heures et 5 jours est-elle valable hors du contexte alpin, voir en Afrique ?

Que penser de la loi log-logistique pour étudier la distribution d'événements extrêmes ? quels sont ses avantages par rapport aux lois utilisées habituellement.

Ceci concerne des problèmes directement liés à la mise en oeuvre du logiciel d'ajustement. Si l'on considère les problèmes indirects liés à l'utilisation des statistiques, beaucoup d'études pourraient être effectuées.

Les lois à plusieurs variables

Il faut tellement d'observations pour vérifier leur adéquation que l'on se limiterait à deux variables (étude des couples : débit de pointe, volume de la crue par exemple).

Un programme concernant la loi normale à 2 variables (et donc log-normale si on travaille sur les logarithmes) existe au L.H.M. sur un support périmé. Ne serait-il pas intéressant de le remettre en forme et de traiter aussi par la même occasion les lois exponentielle et gamma à 2 variables ?

La régionalisation

Il est évident que les pluies ou les écoulements d'une région ont des lois de distribution voisines. N'a-t-on pas créé des Bassins Versants Représentatifs ? On peut compenser en partie la faible ancienneté des observations par leur extension spatiale. Quelles méthodes utiliser ? Quelle est l'amélioration apportée dans l'estimation du risque par rapport à une étude ponctuelle isolée ?

Par ailleurs l'ajustement de lois de probabilité n'est qu'une étape dans un certain nombre de traitements hydrologiques qui font intervenir des processus stochastiques divers.

Les séries chronologiques

Sur une station, les observations de pluie ou de débit constituent une série chronologique ou chronique. Quels sont les processus stochastiques les mieux adaptés à leur modélisation.

L'utilisation des modèles ARMA, ARIMA voir SARIMA des logiciels américains n'est vraiment pas la solution.

Comment traiter ensuite la modélisation d'un ensemble de chroniques à tout un réseau de stations.

La désagrégation

Les hauteurs de pluie journalières sont mesurées depuis longtemps (1767 à Montpellier) ; on a donc une bonne idée de leur distribution.

Les hauteurs de pluie sur des intervalles courts de une heure ou moins sont mesurées depuis peu de temps par des pluviographes.

Comment construire des hyétogrammes horaires qui respectent la distribution des pluies journalières ?

13. CONCLUSION

L'objet de la mission d'évaluation qui nous a été confiée n'était pas de comparer les quatre logiciels étudiés pour décerner des prix de "bonne ou de mauvaise conduite". Les objectifs pour lesquels ils ont été conçus sont très différents, les moyens mis en oeuvre pour les développer également.

ALED est un outil à vocation pédagogique réalisé par des élèves-ingénieurs en Sciences de l'Eau,

DIXLOI a répondu à la volonté de mettre à la disposition des chercheurs hydrologues de l'ORSTOM une aide au traitement statistique.

HFA est un produit conçu par une équipe d'hydrologues, de statisticiens, de mathématiciens-numériciens et d'informaticiens, et donc destiné à une large diffusion.

TROPHEE a été développé à la demande de la Direction Départementale de l'Equipement de l'Ile de la Réunion dans un cadre bien déterminé.

Notre intérêt s'est porté sur la façon dont ces logiciels abordent et traitent chacune des étapes de l'ajustement de lois statistiques sur des variables hydrologiques. Nous avons orienté notre analyse du point de vue du chercheur hydrologue plus ou moins initié à la pratique statistique, et confronté à des questions parfois très complexes.

Nous avons donc examiné un certain nombre de problèmes généraux liés à l'ajustement de lois de probabilité théoriques à un échantillon de données, et plus globalement, à l'estimation des risques en hydrologie.

Pour chaque problème soulevé, nous avons pris des positions étayées par des raisonnements statistiques et/ou hydrologiques ou par des références bibliographiques. Ces positions sont assorties de propositions qui examinées par un groupe de travail comprenant des statisticiens, des informaticiens et des utilisateurs, devraient déboucher sur un cahier des charges définissant un nouveau produit.

BIBLIOGRAPHIE

AHMAD (M.I.), SINCLAIR (C.D.), WERRITTY (A.) 1988
Log-logistic flood frequency analysis. Journal of Hydrology, vol. 98, pp. 205-224.

BOBEE (B.), ASHKAR (F.) 1991
The Gamma family and derived distributions applied in hydrology. Water Resources Publications. 203 pages.

BRUNET-MORET (Y.) 1969
Etude de quelques lois statistiques utilisées en Hydrologie. Cahier ORSTOM, série hydrologie VI, n°3, 100 pages.

BRUNET-MORET (Y.) 1973
Statistiques de rangs. Cahier ORSTOM, série hydrologie, vol. X, n°2, pp. 133-151.

BRUNET-MORET (Y.) 1978
Recherche d'un test d'ajustement. Cahier ORSTOM, série hydrologie, vol. XV, n°3, pp. 261-280.

C.E.R.E.S.T.A 1986
Aide mémoire pratique des techniques statistiques pour ingénieurs et techniciens supérieurs. Numéro spécial de la Revue de Statistique appliquée. Vol. XXXIV, 274 pages.

CRAMER (H.) 1946
Mathematical Methods of Statistics. Princeton University Press. 368 pages.

CUNNANE (C.) 1978
Unbiased plotting positions. A review. Journal of Hydrology, vol. 37, pp. 205-222.

DAVIS (C.S.), STEPHENS (M.A.) 1989
Empirical distribution function goodness-of-fit tests. Algorithm AS 248. Applied Statistics, vol. 38, n°3, pp. 535-543.

DELAFOSSE (E.) 1989
Logiciel "ALED". Avant-projet ISIM-STE (Institut des Sciences de l'Ingénieur de Montpellier, Sciences et Technologies de l'Eau). 40 pages.

DUBAND (D.) 1982
Hydrologie statistique approfondie. Cours donné à l'ENSHG-INPG.

FIORENTINO (M.), GABRIELE (S.) 1984
A correction for the bias of maximum likelihood estimators of Gumbel parameters. Journal of Hydrology, vol. 73, pp. 39-49.

GREENWOOD (J.A.), LANDWEHR (J.M.), MATALAS (N.G.), WALLIS (J.R.) 1979
Probability weighted moments : Definition and relation to parameters of several distributions expressable in inverse form. Water Resources Research, vol. 15, n°5, pp.1049-1054.

GRUBBS (F.), BECK (G.) 1972
Extension of sample size and percentage points for significance tests of outlying observations. Technometrics vol. 14, n°4, pp. 847-854.

HAAN (C.T.) 1977
Statistical Methods in Hydrology. Iowa State University Press/Ames. 378 pages.

HALPHEN (E.) 1949
Les lois des débits des rivières Françaises. La Houille Blanche numéro spécial B. 1949.

HOSKING (J. R. M.), WALLIS (J. R.), WOOD (E. F.) 1985
Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics, vol. 27, n°3, 251-261.

HOUGHTON (J.C.) 1978
Birth of a parent : The Wakeby distribution for modelling flood flows. Water Resources Research, vol. 14, n°6, pp. 1105-1109.

HUYNH (N.P.) 1986
Sampling properties of the maximum entropy estimators for the extreme-value type 1 distribution. Journal of Hydrology, vol. 86, pp. 391-398.

HUYNH (N.P.) 1987
A review of methods of parameter estimation for the extreme value type 1 distribution. Journal of Hydrology, vol. 90, pp. 251-268.

JING (D.), SONG (D.), YANG (R.), HOU (Y.) 1989
Expressions relating probability weighted moments to parameters of several distributions inexpressible in inverse form. Journal of Hydrology, vol. 101, pp. 259-270.

JOWITT (P.W.) 1979
The extreme-value type 1 distribution and the principle of maximum entropy. Journal of Hydrology, vol. 42, pp. 23-38.

KENDALL (S.M.), STUART (A.) 1943
The advanced theory of statistics. Charles Griffin Londres - 3 volumes, 472 pages, 723 pages, 585 pages dans l'édition de 1977.

KISHORE (A.), VIJAY (P.S.) 1989
A comparative evaluation of the estimators of the Log Pearson Type (LP) 3 Distribution. Journal of Hydrology, vol. 105, pp. 19-37.

KITE (G.W.) 1975
Confidence limits for design events. Water Resources Research, vol. 11, n°1, pp. 48-53.

KITE (G.W.) 1976
Frequency and risk analyses in hydrology. Inland waters directorate, water resources branch, Ottawa, Canada. 407 pages.

KUMARASWAMY (P.) 1980
A generalized probability density function for double-bounded random processes. Journal of Hydrology, vol. 46, pp. 79-88.

LABORDE (J.P.) 1984
Analyse des données et cartographie automatique en hydrologie : éléments d'hydrologie lorraine. Thèse d'état INPL. 484 pages.

LEBART (L.), FENELON (J.P.) 1975
Statistique et informatique appliquées. 3ème édition Dunod. 439 pages.

LEBEL (T.), BOYER (J.F.) 1989
DIXLOI : Un ensemble de programmes FORTRAN 77 pour l'ajustement de lois statistiques et leur représentation graphique. Notice OVNIh n°3 du laboratoire d'hydrologie, ORSTOM. 56 pages.

LUBES (H.), MASSON (J.M.) 1991
Méthodes des moments de probabilité pondérés. Application à la loi de Jenkinson. Hydrologie Continentale, à paraître.

LUBES (H.) 1992
Application de lois tronquées aux distributions de précipitations journalières. Actes de SEMINFOR 5 "Statistique Impliquée". 2-3-4 Septembre 1991, Montpellier, Colloques et Séminaires, ORSTOM, Paris, 14 p., sous presse.

MASSON (J.M.) s.d.
La loi des fuites. Document pédagogique ISIM-STE, 21 pages.

MASSON (J.M.) s.d.
La loi de Weibull. Document pédagogique ISIM-STE, 11 pages.

MASSON (J.M.) 1982
La loi gamma incomplète. Document pédagogique ISIM-STE, 36 pages.

MASSON (J.M.) 1983
Méthode générale approchée pour calculer l'intervalle de confiance d'un quantile. Application à quelques lois de probabilité utilisées en Hydrologie. Document pédagogique ISIM-STE, 41 pages.

MASSON (J.M.) 1983
La loi de Gumbel. Document pédagogique ISIM-STE, 40 pages.

MASSON (J.M.) 1985
La loi log-normale. Document pédagogique ISIM-STE, 32 pages.

MASSON (J.M.) 1992
Un problème parmi d'autres dans l'analyse des distributions des variables hydrologiques : les horsains (outliers). Actes de SEMINFOR 5 "Statistique Impliquée". 2-3-4 Septembre 1991, Montpellier, Colloques et Séminaires, ORSTOM, Paris, 10 p., sous presse.

MERCIER (B.) 1991
Ajustement des principales lois statistiques utilisées en hydrologie. Rapport d'ingéniorat ISIM-STE, 56 pages.

MICHEL (C.) 1989
Hydrologie appliquée aux petits bassins ruraux. CEMAGREF, Antony. 528 pages.

Natural Environment Research Council (N.E.R.C) 1975
Flood studies Report. Hydrological Studies, vol. I, 549 pages.

RAYNAL (J.A.), SALAS (J.D.) 1986
Estimation procedures for the type-1 extreme value distribution. Journal of Hydrology, vol. 87, pp. 315-336.

RISSONS (M.) 1988
Logiciel d'hydrologie statistique. Avant-projet ISIM. 3 parties de 47, 45 et 58 pages.

ROCHE (M.) 1963
Hydrologie de surface. Gauthier-Villars - Paris 430 p.

ROSSI (F.), FIORENTINO (M.), VERSACE (F.P.) 1984
Two Component extreme value distribution for flood frequency analysis.
Water Resources Research, vol. 20, n°7, pp. 847-856.

SCOTT (D.W.) 1985
Frequency Polygons : Theory and Application Journal of the American Statistical Association, vol. 80, n° 390, pp. 348 à 353.

SONG (D.), DING (J.) 1988
The application of probability weighted moments in estimating the parameters of the Pearson type three distribution. Journal of Hydrology, vol. 101, pp. 47-63.

WALD (A.), WOLFOWITZ (J.) 1943
An exact test for randomness in the non parametric case based on serial correlation. Ann. math. statist., 14, pp. 378-388.

WEIBULL (W.) 1951
Statistical distribution function of wide application. Journal of Applied Mechanics ASME. Vol. 18 - p. 293-297.

WHITE (J.C.) 1964
Least square unbiased censored linear estimation for the Log Weibull (extreme value) distribution. J. Ind Math., 14, pp. 21-60.